Continuous respiratory rate monitoring to detect clinical deteriorations using wearable sensors

Charlton, Peter

Awarding institution:
King's College London

Continuous respiratory rate monitoring to detect clinical deteriorations using wearable sensors

Charlton, Peter

Awarding institution:
King's College London

Continuous Respiratory Rate Monitoring to Detect Clinical Deteriorations using Wearable Sensors

Peter H. Charlton

Supervised by

Dr. Jordi Alastruey

Prof. Richard Beale

Prof. David Clifton

June 2017

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the

Division of Imaging Sciences and Biomedical Engineering

Faculty of Life Sciences and Medicine

This thesis is dedicated to three patients:

A patient who tolerated a wearable sensor poorly, until it detected that they were having a cardiac arrest and arguably dramatically improved their future health. Thereafter they tolerated the sensor very well.

A patient who was very talkative before their operation, and died tragically young a few days later despite wearing a wearable sensor.

A patient who probably didn't need a wearable sensor, but ended up practically dancing around the ward whilst wearing it, laughing and joking with the staff.

' . . . a time to mourn and a time to dance . . . '

Ecclesiastes, ch. 3

Abstract

Acutely-ill hospitalised patients are at risk of clinical deteriorations in health leading to adverse events such as cardiac arrests. Deteriorations are currently detected by manually measuring physiological parameters every 4-6 hours. Consequently, deteriorations can remain unrecognised between assessments, delaying clinical intervention. It may be possible to provide earlier detection of deteriorations by using wearable sensors for continuous physiological monitoring. Respiratory rate (RR) is not commonly monitored by wearable sensors, despite being a sensitive marker of deteriorations. This thesis presents investigations to identify an algorithm suitable for estimating RR from two signals commonly acquired by wearable sensors: the electrocardiogram (ECG) and photoplethysmogram (PPG). A suitable algorithm was then used to estimate RRs retrospectively from a physiological dataset acquired from acutely-ill patients to assess the potential utility of wearable sensors for detecting deteriorations.

Existing RR algorithms were identified through a systematic review of the literature. A toolbox of RR algorithms was created to facilitate comprehensive assessments of algorithms across multiple datasets. This was used to assess the influence of technical and physiological factors on respiratory signals extracted from the ECG and PPG, providing recommendations for wearable sensor designs for RR estimation. An assessment of 95 RR algorithms using data from healthy and hospitalised patients showed that the algorithms did not perform well enough for use with acutely-ill patients. Therefore, a novel algorithm was designed specifically for use with wearable sensors, providing improved performance.

The novel RR algorithm was used to estimate RRs retrospectively from wearable sensor data acquired from 184 patients. The performances of algorithms to detect deteriorations from the resulting wearable sensor data were similar to those used with routinely collected intermittent data, suggesting that it is feasible to use wearable sensors to continuously assess the likelihood of deterioration. However, the false alert rate increased when using wearable sensor data due to the continuous, rather than intermittent, monitoring. Therefore, further work is required to improve algorithms to detect deteriorations from wearable sensor data to provide clinically useful alerts.

Licence

Acknowledgements

I am grateful to the many people who have supported and encouraged me throughout this work.

Firstly, I'd like to thank my supervisors: Jordi Alastruey, Richard Beale and David Clifton. Jordi has been a great role model, has constantly provided valuable advice, and has taught me much about the measurement of cardiovascular physiology. Richard has given me much motivation for pursuing this line of research, through the example of his own determination and leadership, and by allowing me to work alongside clinicians and patients in St Thomas' Hospital. David has provided guidance on biomedical signal processing and machine learning techniques, and has been a source of encouragement and inspiration throughout.

Secondly, I'd like to thank the Critical Care Department at Guy's and St Thomas' NHS Foundation Trust. It has been a pleasure and a privilege to work alongside the dedicated clinicians within the department, and they have provided an essential platform without which this work would not have been possible. In particular, I'd like to thank the following: John Smith, Isabelle Schelcher, Ricky Yang, John Brooks, Katie Lei, Luigi Camporota, Marius Terblanche, Marlies Ostermann, Tony Sherry, and Barney Sanderson.

Thirdly, I'd like to thank my collaborators. I've very much enjoyed working with Marco Pimentel, Christina Orphanidou, David Vallance, Mauro Villaroel, Peter Watkinson, Philip Aston, Manasi Nandi, Mark Christie and Leo Celi. My thanks also to those who have funded this work, namely the UK Engineering and Physical Sciences Research Council, the National Institute for Health Research (NIHR), and King's College London.

I'm also grateful to family and friends who have been a great support. Sarah, thanks for always being beside me and for your unwavering love. Mum, Dad, Mia and Sue, thanks for all your encouragement through the years. Dad is, without a doubt, the single academic I look up to the most. Ant, thanks for your friendship throughout (and for the encouragement [1]). Thank you to those at Rye Lane Baptist Chapel and All Saints Peckham, especially Frank, Rhona, Barry and Gwen. Thank you too to those who have been a support through their prayer and friendship: Dave, Tom, John, Ben, Pete, Dan, Greg.

Finally, I'd like to thank Lionel Tarassenko for giving me the inspiration and the opportunity to work in this field. Many thanks to Tim Bonnici, whose hard work and perseverance has made this research not only possible, but also a lot of fun. I'd also like to thank the patients who gave up their time and energy, without whom this work would have been not only impossible, but also fairly meaningless.

Peter Charlton, June 2017

Related Publications

This thesis includes work that has also been reported in the following related publications.

I presented the processing of the LISTEN dataset described in Chapter 3 at the following conferences:

[2] **P. Charlton** and T. Bonnici, D. A. Clifton, L. Tarassenko, P. Watkinson and R. Beale, "Achieving Clinical Quality from Wireless Sensors," *Conf IEEE Eng Med Biol Soc.* Chicago, IL. 2014

[3] T. Bonnici and **P. Charlton**, J. Alastruey, L. Tarassenko, P. J. Watkinson and R. Beale, "Continuous Physiological Monitoring of Ambulatory Patients," in *MEC Annual Meeting and Bioengineering14 Programme and Abstracts.* London: MECbioeng14, Imperial College London, 2014. p. 38. **CrossRef**

The structure of respiratory rate (RR) algorithms identified through the literature review presented in Chapter 4 was first reported in:

[4] M. A. F. Pimentel, **P. H. Charlton** and D. A. Clifton, "Probabilistic estimation of respiratory rate from wearable sensors," in *Wearable Electronics Sensors*, Springer International Publishing, 2016. pp. 241 - 62. **CrossRef**

The literature review has been expanded in:

P. H. Charlton, D. A. Birrenkott, T. Bonnici, M. A. F. Pimentel, A. E. W. Johnson, J. Alastruey, L. Tarassenko, P. J. Watkinson, R. Beale and D. A. Clifton, "Respiratory rate estimation from the electrocardiogram and photoplethysmogram: a review," [under review].

In addition, the RR estimation resources described in Chapter 4 are being made publicly available at the **RRest Project** website. They were presented at the Science and Engineering South Data Dialogue Meeting [5] (**link**). The toolbox of algorithms was also introduced in this tutorial:

[6] **P. H. Charlton**, M. Villarroel and F. Salguiero, "Waveform analysis to estimate respiratory rate," in *Secondary Analysis of Electronic Health Records*, Springer International Publishing, 2016. pp. 377 - 90. **CrossRef**

The analysis of factors influencing the quality of respiratory signals extracted from the ECG and PPG, described in Chapter 5, has been published in:

[7] **P. H. Charlton**, T. Bonnici, L. Tarassenko, J. Alastruey, D. A. Clifton, R. Beale and P. J. Watkinson, "Extraction of Respiratory Signals from the Electrocardiogram and Photoplethysmogram: Technical and Physiological Determinants," *Physiological Measurement*, 38(5), pp. 669 - 90, 2017. **CrossRef**

This work was also presented at the MECbioeng14 conference:

[8] **P. Charlton** and T. Bonnici, D. Clifton, J. Alastruey, L. Tarassenko, P. J. Watkinson and R. Beale, "The influence of recording equipment on the accuracy of respiratory rate estimation from the electrocardiogram and photoplethysmogram," in *MEC Annual Meeting and Bioengineering14 Programme and Abstracts*. London: MECbioeng14, Imperial College London, 2014. p. 96. **CrossRef**

Results from the assessment of respiratory rate algorithms described in Chapter 6 were presented at the MEIbioeng 16 conference (University of Oxford) [9] (**link**), and published here:

[10] **P. H. Charlton** and T. Bonnici, L. Tarassenko, D. A. Clifton, R. Beale and P. J. Watkinson, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiological Measurement*, 37(4), pp. 610 - 626, 2016. **CrossRef**

This paper has been awarded the Martin Black Prize for best paper published in Physiological Measurement in 2016.

Initial work on the development of algorithms to detect deteriorations was published here:

[11] **P. H. Charlton**, M. Pimentel and S. Lokhandwala, "Data fusion techniques for early warning of clinical deterioration," in *Secondary Analysis of Electronic Health Records*, Springer International Publishing, 2016. pp. 325 - 38. **CrossRef**

Contents

List of Figures

List of Tables

Abbreviations

ABP	**A**rterial **B**lood **P**ressure
AE	**A**dverse **E**vent
AFib	**A**trial **Fib**rillation
AM	**A**mplitude **M**odulation
AR	**A**uto**r**egressive
AUROC	**A**rea **U**nder **R**eceiver **O**perator **C**urve
bpm	**b**reaths **p**er **m**inute, *or* **b**eats **p**er **m**inute
BW	**B**aseline **W**ander
C	**C**ost function
CC	**C**orrelation **C**oefficient, *or* **C**ritical **C**are
CVP	**C**entral **V**enous **P**ressure
CABG	**C**oronary **A**rtery **B**ypass **G**raft
CP	**C**overage **P**robability
DBP	**D**iastolic **B**lood **P**ressure
ECG	**E**lectro**c**ardio**g**ram
eCRF	**e**lectronic **C**ase **R**eport **F**orm
EHR	**E**lectronic **H**ealth **R**ecord
EMG	**E**lectro**m**yo**g**ram
EWS	**E**arly **W**arning **S**core
FM	**F**requency **M**odulation
HR	**H**eart **R**ate
HRV	**H**eart **R**ate **V**ariability
ICU	**I**ntensive **C**are **U**nit
iCP	**i**nverse **C**overage **P**robability
ImP	**Im**pedance **P**neumography
InP	**In**ductance **P**lethysmography
KDE	**K**ernel **D**ensity **E**stimation
LoAs	**L**imits **o**f **A**greement
LOC	**L**evel **O**f **C**onsciousness
MAP	**M**ean **A**rterial **P**ressure
NEWS	**N**ational **E**arly **W**arning **S**core
O_2	**O**xygen therapy

PDF	**P**robability **D**ensity **F**unction
PP	**P**ulse **P**ressure
PPG	**P**hoto**p**lethysmo**g**ram
PR	**P**ulse **R**ate
prop	**p**roportion of windows
PPV	**P**ositive **P**redictive **V**alue
PTT	**P**ulse **T**ransit **T**ime
RQI	**R**espiratory **Q**uality **I**ndex
RR	**R**espiratory **R**ate
RSA	**R**espiratory **S**inus **A**rrhythmia
SAE	**S**evere **A**dverse **E**vent
SBP	**S**ystolic **B**lood **P**ressure
SpO_2	Arterial Blood Oxygen Saturation
SQI	**S**ignal **Q**uality **I**ndex
temp	**temp**erature
VHF	**V**ery **H**igh **F**requency
VLF	**V**ery **L**ow **F**requency

Chapter 1

Introduction

Acutely-ill hospitalised patients are at risk of deteriorations in health leading to adverse events such as cardiac arrests, sepsis, and acute kidney injury. Adverse events such as these are detrimental to patients as they increase morbidity and mortality, and prolong their length of stay in hospital [12–14]. In addition, adverse events have a significant impact on healthcare costs and resources. For instance, in the USA adverse events have been estimated to cost upwards of US$17 billion per annum [15], and in Australia they have been estimated to account for 8% of hospital bed days [16]. Furthermore, in the UK clinical negligence costs for the National Health Service (NHS) totalled £769 million during 2008-2009 [17]. Therefore, any measures which reduce the incidence of adverse events may improve patient outcomes and reduce healthcare resource utilisation.

Many adverse events are preceded by measurable changes in physiology. Schein *et al.* published landmark results in 1990 that 84% of patients *"had documented observations of clinical deterioration or new complaints"* in the eight hours preceding cardiac arrest [18]. This finding was subsequently supported by a study by Franklin *et al.* [19]. Physiological abnormalities have also been observed in the hours prior to unplanned admissions to the Intensive Care Unit (ICU) [20, 21] and preventable deaths [22]. Furthermore, in a study of electronic health record (EHR) variables, it was found that 9 out of 14 predictors of cardiopulmonary arrests, unplanned transfers to ICU, and death, were physiological (and at least 4 of the remaining 5 were directly linked to physiology) [23]. The assumption that adverse events are preceded by changes in physiology is also supported by expert opinion. In 2010 a consensus of international experts found that most cardiac arrests and deaths are preceded by physiological parameters *"lying*

outside . . . normal ranges" [24]. The overwhelming conclusion is that changes in physiology can be observed during the hours prior to many adverse events.

It is widely accepted that the progression of deteriorations, and therefore the likelihood of subsequent adverse events, can be reduced by recognising and acting on the accompanying changes in physiology [25–27]. Currently, staff on general hospital wards identify deteriorations through manual measurement of physiological parameters every 4-6 hours (such as heart rate and blood pressure), from which an early warning score (EWS) is calculated indicating the likelihood and severity of a deterioration [28]. The effectiveness of current practice is limited to intermittent assessments since they have to be performed manually. This means deteriorations can remain unrecognised until the next assessment, potentially delaying the clinical response by several hours. In addition, the current manual assessments are time-consuming and inefficient.

It may be possible to address the shortcomings of manual and intermittent measurements by using wearable sensors to monitor physiology continuously and automatically. Wearable sensors for use in hospitals have traditionally been limited to providing electrocardiogram (ECG) based heart rate and arrhythmia monitoring. Consequently their use has been confined to patients at risk of arrhythmias. Recent developments in wearable sensor design and biomedical signal processing techniques mean that it may now be possible to monitor a broader range of physiological parameters using wearable sensors [29]. As a result, wearable sensors may now have utility for the early detection of the wider range of deteriorations from which all acutely-ill inpatients are at risk [30]. For example, wearable sensors can now monitor additional signals such as the pulse oximetry (photoplethysmogram, PPG) signal, providing arterial blood oxygen saturation (SpO_2) monitoring. Furthermore, signal processing techniques for monitoring of the respiratory, vascular, and autonomic systems have been developed in other settings, which could be adapted for use with wearable sensors. Consequently, novel wearable sensors may be useful for detecting not only cardiac deteriorations, but also those which manifest as changes in other physiological systems.

Two additional steps are required to realise the potential of wearable sensors to provide clinical benefit through detection of deteriorations. The first step concerns respiratory rate (RR), which is a highly sensitive marker of deterioration [31]. Current methods for monitoring RR using wearable sensors are either obtrusive or inaccurate [32]. Therefore, further development and evaluation of RR monitoring techniques for use in wearable sensors is required. Secondly, techniques for generating alerts of deteriorations need to be adapted for use with wearable

sensors. Wearable sensor data is susceptible to artifact, due to poor sensor contact and movement, and is also acquired during periods of activity. In contrast, intermittent physiological measurements are checked for artifact by the measurer, and are acquired whilst patients are at rest. Therefore, the techniques which have previously been used to identify deteriorations from intermittent measurements may need to be refined for use with wearable sensor data. If these issues can be resolved then there is potential for wearable sensors to improve patient safety and quality of care, and reduce staff workload [33]. This is of particular importance since the number of patients in hospital and severities of illnesses are increasing, whilst staffing levels are decreasing [29, 34].

1.1 Thesis Goals

This thesis addresses the first step identified to realise the potential of wearable sensors: developing RR monitoring techniques for use in wearable sensors. The overall aim of this thesis is:

> To develop and assess the performance of techniques for continuous RR monitoring using ECG and PPG signals for use in wearable sensors to detect deteriorations.

A deterioration is defined as an acute worsening of health, resulting in a transition to a physiological state with an increased likelihood of a clinical adverse event [35].

Several goals were identified in order to achieve the overall aim as follows:

- **To assemble a comprehensive physiological dataset using wearable sensors:** A comprehensive dataset was assembled containing wearable sensor data and labels of deteriorations suitable for development of algorithms to detect deteriorations. This dataset included continuous ECG and PPG signals, from which RR could be estimated retrospectively.
- **To identify algorithms to estimate RR from the ECG and PPG from the literature, and methods used to assess their performance, to create a toolbox of resources for algorithm assessments:** Many algorithms have been proposed to estimate RR from the ECG and PPG signals, which are commonly acquired by wearable

sensors. These provide a convenient method for monitoring RR. However, a systematic review of the literature found that these algorithms had not been comprehensively assessed. Therefore, it was unclear which algorithm, if any, performed well enough for use in wearable sensors. A toolbox of algorithms and benchmark datasets was created to facilitate comprehensive assessments of algorithm performances.

- **To assess the influence of technical and physiological factors on respiratory signals extracted from the ECG and PPG:** It was not clear how technical and physiological factors influence the performance of RR algorithms. Therefore, the quality of respiratory signals extracted from the ECG and PPG was assessed under different signal acquisition procedures and a range of physiological conditions to provide recommendations for the design of wearable sensors incorporating RR algorithms, and for their clinical use.

- **To assess the performance of RR algorithms for the ECG and PPG in healthy subjects and hospitalised patients:** A comprehensive assessment of existing RR algorithms was performed to determine which, if any, were suitable for continuous monitoring of acutely-ill patients using wearable sensors.

- **To develop a novel algorithm for continuous RR monitoring:** A novel RR algorithm was designed specifically for use with acutely-ill patients and wearable sensors. This algorithm was found to provide superior performance compared to existing RR algorithms.

- **To assess the potential utility of wearable sensors for detecting deteriorations:** The potential utility of continuous RR monitoring using wearable sensors to detect deteriorations was assessed by using the novel RR algorithm to estimate RRs retrospectively from the comprehensive physiological dataset assembled in the first goal. This provided evidence to suggest that EWSs could be continuously updated using data acquired from wearable sensors.

1.2 Thesis Outline

This thesis is structured as follows. Chapter 2 presents the clinical background, including the current clinical practice for identifying deteriorations in ambulatory hospital patients, the growing role of wearable sensors, and the steps required to realise the potential of wearable sensors to improve identification of deteriorations. In Chapter 3 the data processing steps

taken to assemble a comprehensive physiological dataset are described. The resulting dataset is suitable for the development of algorithms to detect deteriorations. In Chapter 4 a systematic review of RR algorithms is presented, followed by a toolbox of resources for assessment of RR algorithms. Chapter 5 reports a study of the influence of technical and physiological factors on the quality of respiratory signals extracted from the ECG and PPG, with implications for the use of RR algorithms in wearable sensors. Chapter 6 presents a study of the performance of RR algorithms across a range of datasets, culminating in the conclusion that no existing RR algorithm performed well enough to be used for continuous RR monitoring of acutely-ill patients. In Chapter 7 a novel RR algorithm is presented, designed specifically for monitoring acutely-ill patients. In Chapter 8 the novel RR algorithm was used with the dataset presented in Chapter 3 to demonstrate the feasibility of continuous RR monitoring to detect clinical deteriorations using wearable sensors. Finally, Chapter 9 presents a summary of the achievements of this thesis, and directions for future work.

Chapter 2

Clinical Background

In this chapter the clinical background to this thesis is presented. Firstly, the clinical need for identification of deteriorations is described. Secondly, an overview of current practice for physiological assessment of ambulatory patients, and identification of deteriorations, is presented. This is followed by an explanation of the potential role of wearable sensors in providing earlier identification of deteriorations than current practice, and reducing staff workload. The current evidence for the clinical benefit of multi-parametric wearable sensors is then reviewed. Finally, the steps required to realise the potential of wearable sensors are described: (i) the need for continuous, unobtrusive, RR monitoring; and, (ii) the need for sensitive and specific algorithms to generate alerts of deteriorations from continuous wearable sensor data.

2.1 The Need for Early Identification of Deteriorations

There is no consensus on the definition of a clinical deterioration in the literature [36]. Deteriorations have been defined using four approaches: (i) adverse events contributed to by negligence; (ii) defined clinical adverse events; (iii) the presence of physiological instability; and (iv) movement to a clinical state which is associated with increased risk of defined clinical adverse events [36]. In this work approach (iv) was used. A clinical deterioration was defined as an acute worsening of health, resulting in transition to a physiological state with an increased likelihood of a clinical adverse event. Examples of clinical adverse events include the development of sepsis, acute kidney failure, respiratory failure, and cardiac arrest. This definition of a deterioration

is particularly relevant to both patients and clinicians, since clinical adverse events result in increased morbidity, mortality, and healthcare resource utilisation.

Acutely-ill hospitalised patients are at increased risk of deteriorations. Within this broad patient population, specific groups have a particularly high risk of deteriorations. For instance, approximately 50 % of patients recovering from upper gastrointestinal surgery suffer from an adverse event during their hospital stay [37]. In contrast, approximately 10 - 20 % of patients recovering from cardiac surgery suffer an adverse event whilst in hospital [38, 39]. A large proportion of adverse events (38.7 - 60.7 % in different patient groups) in post-operative patients are potentially avoidable [39]. Therefore, there is a clinical need to identify deteriorations early to prevent the subsequent adverse events which increase morbidity, mortality, and healthcare costs.

Acutely-ill hospital patients suffer from a wide range of different types of adverse events [39, 40]. The relative importance of detecting deteriorations which are associated with each type of adverse event is determined by several factors. Firstly, different types of adverse events have different incidence rates in different populations. For instance, in a study of patients recovering from coronary artery bypass graft (CABG) surgery, approximately 9 % suffered from congestive heart failure, whereas only 1 % suffered from a stroke [40]. Secondly, the impact of each type of adverse event on morbidity, mortality and quality of life varies [41]. For example, in this CABG study the mortality rates associated with the two types of adverse events differed markedly, with approximately 12 % of occurrences of congestive heart failure resulting in death, compared to 27 % of strokes. Thirdly, different types of deteriorations can be treated with differing success rates. In this example, the treatments for the two deteriorations differed. The symptoms of congestive heart failure can be treated by oxygen (or respiratory) therapy, and the use of vasodilators and diuretics. In contrast, the treatments for strokes are less easily administered, and often last far longer, including extensive rehabilitation programmes. The heterogeneity in incidence, impact, and treatments of different types of deteriorations has an effect on the relative importance of detecting each one.

It is important that deteriorations are identified early to improve patient outcomes and reduce healthcare costs. In a study of inpatients from four surgical services approximately 6 - 12 % of deteriorations resulted in death [39]. Furthermore, 20 - 45 % of these deaths were judged to be potentially avoidable [39]. Therefore, timely identification and appropriate management of deteriorations may decrease mortality. The importance of early identification of deteriorations

was shown by a trial investigating the impact of hospital-wide medical emergency teams (METs) [42], who attend to and manage deteriorating patients. The rate of cardiac arrests and unexpected deaths decreased as the proportion of early MET calls increased [43], suggesting that early identification and management of deteriorating patients may indeed be associated with improved outcomes. In addition, deteriorations have been shown to increase length of stay after surgery [12], thereby increasing healthcare costs [13, 14]. Early identification of deteriorations is beneficial for both patients and healthcare providers.

2.2 Physiological Monitoring of Ambulatory Inpatients

Acutely-ill hospital patients have their physiology measured regularly to assess their risk of deterioration, since deteriorations are commonly accompanied by changes in physiology [18],[19],[20–24]. In the UK seven physiological parameters are measured regularly in acutely-ill patients, as listed in Table 2.1 [28]. These physiological parameters are typically measured every 4-6 hours, although this can be decreased to once every 12 hours for patients not displaying any deranged physiological parameters, or increased to every hour or even continuous monitoring for patients displaying deranged physiology [28]. Physiological measurements are obtained using a monitor which is typically wheeled from one patient to the next. The current practice of taking regular physiological measurements is designed to ensure that every acutely-ill patient is regularly assessed for signs of deterioration, and has several benefits:

- **Accuracy:** Since measurements are taken by hand, the operator can assess the accuracy of the measurements, quickly resolving the causes of any unreliable measurements, such as poor sensor contact.
- **Data coverage:** The level of patient compliance with these physiological measurements is extremely high, since an operator ensures that the measurements are successfully taken.
- **Additional assessment:** The operator has opportunity to observe additional factors which may be indicative of deterioration, which cannot be measured electronically.
- **Response:** A clinical response can be easily initiated in response to the physiological measurements, because these are immediately reviewed by the operator. The operator can either initiate a response themselves (such as the administration of supplemental oxygen),

TABLE 2.1: Physiological parameters measured intermittently during routine practice: The listed parameters are measured every 4 - 6 hours in acutely-ill hospital patients to calculate the National Early Warning Score (NEWS) [28]. Details of techniques (either electronic or manual) which are commonly used to measure the parameters are provided.

Parameter	Abbr.	Measurement technique	Electronic?
Heart rate	HR	A pulse oximeter is used to automatically estimate HR from the PPG.	✓
Arterial blood oxygen saturation	SpO_2	A pulse oximeter is used to automatically estimate SpO_2 at the same time as estimating HR.	✓
Systolic blood pressure	SBP	A sphygmomanometer is used to automatically estimate SBP. This involves application of a blood pressure cuff around the upper arm.	✓
Temperature	temp	A thermometer is used to estimate core temperature. A variety of thermometers are in routine use, including single-use oral thermometers, and electronic tympanic thermometers.	✓
Respiratory rate	RR	RR (the number of breaths per minute) is usually measured by manually counting chest wall movements over a period of approximately 30 s.	✗
Oxygen therapy	O_2	A binary variable indicating whether or not the patient is receiving supplementary oxygen.	✗
Level of consciousness	LOC	A multinomial variable indicating the LOC with the following categories: (i) Alert (fully awake), (ii) Voice (responds to a voice stimulus), (iii) Pain (responds to a pain stimulus), (iv) Unresponsive (no response to voice or pain).	✗

or can quickly inform another member of staff to trigger a more complex response (such as a new pharmacological intervention).

Once a set of measurements have been obtained an additional step is taken to identify deteriorations. This is performed by calculating an early warning score (EWS), an aggregate score calculated from physiological parameters, which indicates the presence and severity of a deterioration. EWSs are designed to overcome the problem that no single parameter is all-important for identification of deteriorations [44]. A simple formula is used in which each parameter is assigned a score indicating its level of abnormality, and these individual scores are summed to give an aggregate score. The higher the aggregate score, the higher the likelihood of deterioration. The regular use of a standardised National Early Warning Score (NEWS) to assess acutely-ill hospital patients was mandated in the UK in 2012 [28]. The methodology for calculating the NEWS is displayed in Table 2.2. This score has been found to outperform previously proposed

TABLE 2.2: Calculation of the National Early Warning Score (NEWS): Seven physiological parameters (defined in Table 2.1) are measured, and a score is assigned to each. The NEWS is then calculated as the sum of these individual scores [28]. Definitions of LOC categories: A - alert, V - responds to voice, P - responds to pain, U - unresponsive.

Parameter	Score assigned to individual parameter						
	3	2	1	0	1	2	3
HR	≤ 40		41-50	51-90	91-110	111-130	≥ 131
SpO_2	≤ 91	92-93	94-95	≥ 96			
SBP	≤ 90	91-100	101-110	111-219			≥ 220
temp	≤ 35.0		35.1-36.0	36.1-38.0	38.1-39.0	≥ 39.0	
RR	≤ 8		9-11	12-20		21-24	≥ 25
O_2		Yes		No			
LOC				A			V, P, or U

EWSs when used to identify patients at risk of cardiac arrest, unexpected ICU admission, and death [45]. EWSs have also been integrated into worldwide clinical practice [46].

The performances of EWSs are the subject of debate. On one hand, EWSs have been shown to correlate with important patient-centred endpoints such as levels of intervention [47], higher hospital mortality [47, 48], length of stay [48], and to be a better predictor of cardiac arrest than individual parameters [49]. Furthermore, they have resulted in clinical response at lower levels of physiological abnormality [50]. On the other hand, it has not been shown conclusively that the use of EWSs translates to improved patient outcomes [51]. For instance, in 2003 a study of the use of an EWS to triage medical admissions did not find any impact on outcomes [31]. In 2007, there was *"little evidence of reliability, validity and utility"* of EWSs [52], particularly as the majority of 33 different EWSs discriminated poorly between survivors and non-survivors [53]. Reassuringly, since then the NEWS has been recommended for use across the NHS, and has been shown to outperform the 33 previously tested EWSs [45]. Nonetheless, there may still be scope to improve the performances of EWSs.

2.3 The Potential Role for Wearable Sensors

The use of wearable sensors (hereafter also defined as wireless) to monitor human physiology began in 1961 when the first human in space, Yuri Gagarin, was continuously monitored by

doctors on earth [29]. By 1969 the technology had developed to allow simultaneous monitoring of multiple astronauts during the first moon-landing mission. By the 1980s wearable sensors had been successfully translated into hospital use in intensive and acute care units [54].

Since the 1980s the role of wearable sensors in hospitals has been largely confined to cardiac monitoring. Their routine use is dominated by electrocardiogram (ECG) monitoring, which is indicated for specific patient groups in hospital [55]. Wireless ECG monitoring is typically used to facilitate mobilisation when patients are not sufficiently ill to require more intensive, tethered monitoring. Possible settings include intensive and coronary care units, cardiac telemetry wards, and less routinely, post-cardiac surgery units, the surgical recovery room, and the emergency department [56].

Wearable sensors are currently used to identify abnormal heart rhythms. A wide range of rhythms can be identified in real-time, such as: inefficient arrhythmias including atrial fibrillation; life-threatening arrhythmias such as bradycardia and ventricular tachycardia; and asystole. Indeed, the use of wearable sensors for identification of life-threatening dysrhythmias has been shown to improve patient outcomes. Recent studies have found telemetry (remote monitoring, typically of the ECG) to be associated with improved survival to hospital discharge following in-hospital cardiac arrest [57, 58]. The mechanism underlying this improvement may be that the use of telemetry allowed earlier defibrillation, which increased the likelihood of survival to hospital discharge [59]. This illustrates how wearable sensors are currently used to trigger restorative, rather than preventative, treatments.

There is a desire to use wearable sensors in a preventative manner for early detection of the deteriorations which precede adverse events. Currently, clinical deteriorations of acutely-ill patients are detected through manual physiological assessments. However, this is limited to intermittent assessments, meaning deteriorations can remain unrecognised until the next assessment exposing the patient to delays in time-sensitive interventions. The use of wearable sensors coupled with an intelligent notification system may facilitate earlier detection of deteriorations, since they could provide continuous physiological assessments. Indeed, in some cases they are currently being used for this purpose, rather than solely for detection of dysrhythmias [60], despite only being able to monitor a minority of physiological parameters such as HR and SpO_2. However, studies have demonstrated only minimal, or no, improvement in patient outcomes associated with the use of wearable sensors for detection of deteriorations [56, 61–63]. This is because traditional wearable sensors are designed to identify dysrhythmias based on ECG

monitoring, but are ill-equipped to identify the multi-parametric physiological changes that indicate deteriorations. Furthermore, there is a low frequency of cardiac arrests or dysrhythmias amongst patients monitored by telemetry [61–63], meaning that wearable sensors will have limited impact unless a wide range of deteriorations can be detected. There are also unintended consequences of using wearable sensors in this way. They may provide a false sense of security [30] and result in alarm fatigue due to a high false alarm rate [64]. Overall, the evidence suggests that the use of traditional ECG-based wearable sensors for detecting deteriorations is far from ideal, as summarised in [65].

Recent developments have rendered wearable sensors much more suitable for detection of deteriorations. Firstly, novel wearable sensors can now be used to monitor a much broader range of physiological parameters than traditional wearable sensors [29]. Table 2.3 lists those parameters which can be measured by wearable sensors which are available for clinical use in hospitals. These include the five routinely measured 'vital signs' (HR, RR, SpO_2, SBP and temp), and ECG-based analyses. This demonstrates that it is technically possible to monitor the five vital signs using a wearable sensor, as well as parameters indicative of cardiac function. Secondly, wearable sensors are being made less obtrusive, making them more acceptable to patients. For example, wrist-worn and patch-style sensors are now being trialled for use in hospitals, such as the sensor shown in Figure 2.1 [66, 67]. Thirdly, algorithms for assessing the quality of physiological signals can now be implemented on wearable sensors to ensure that only parameters estimated from high quality signals are reported (such as HRs estimated from the ECG) [67]. Fourthly, the battery life of wearable sensors has increased up to five days [68], reducing the workload associated with battery replacement. Fifthly, wearable sensors can now be operated on a hospital's Wi-Fi network, rather than requiring a bespoke wireless network [69]. Finally, novel data fusion algorithms for identification of deteriorations from multi-parametric data have been developed [70], increasing the reliability of identifications of deteriorations [71]. As a result, the role of wearable sensors may no longer be largely confined to cardiac monitoring in patients perceived to be at risk of dysrhythmias [30]. Instead, they may have utility for the early detection of deteriorations which precede adverse events.

An additional, related application for wearable sensors is to reduce the workload associated with manual physiological assessments. Considerable workload is created by 4 - 6 hourly assessments since on a general hospital ward there are several patients per member of staff, and each manual assessment takes several minutes. It may be possible to use wearable sensors to safely reduce the frequency of manual assessments in patients deemed to be at low risk of deterioration. A process

TABLE 2.3: Parameters which can be measured using wearable sensors

Parameter	Sensor(s)	Methodology
		Vital Signs
HR	ECG or pulse oximeter	Heart rate (HR) can be measured from either the ECG or PPG by detecting individual heart beats in the signals. This is commonly achieved by using a QRS detector (*e.g.* [72, 73]), or pulse peak detector (*e.g.* [74]).
SpO_2	pulse oximeter	Arterial blood oxygen saturation (SpO_2) can be measured using a pulse oximeter, commonly situated at either the finger or ear. The acceptability of pulse oximeters is improving with the design of more comfortable sensors, such as a miniaturised Velcro finger strap sensor [75].
RR	ImP, InP or accelerometer	Several techniques have been developed for monitoring respiratory rate (RR), although none is used ubiquitously. RR can be extracted from: inductance plethysmography (InP) signals acquired from a strain gauge in a chest band [76]; impedance pneumography (ImP) signals acquired using two ECG electrodes [69]; and recently an accelerometer has been used to monitor respiratory movements on the abdomen [77]. Although not in clinical use, methods are being developed for estimation of RR from a gyroscope, the ECG, and the PPG [78–80].
BP	sphygmomanometer	Some novel wearable sensors include a sphygmomanometer for blood pressure (BP) measurement [77].
temp	thermometer	Surface temperature (temp) can be measured in wearable sensors using a thermometer sited at various locations including the axilla [81] or at an ECG electrode [69].
		ECG-Based Analyses
Arrhythmia Detection	ECG	Arrhythmia and dysrhythmia alarms are routinely provided in traditional wearable sensors through analysis of the ECG [82].
ECG morphology	ECG	Novel wearable sensors provide measurements of ECG morphology such as ST duration and QT elevation [83].

such as that illustrated in Figure 2.2 could be used to achieve this. In this process, patients who have a low enough NEWS at a manual assessment would have the next manual assessment conducted 12 hours later unless their wearable sensor detected a change in physiology. If a change in physiology was detected by the wearable sensor, then a manual assessment would be conducted immediately. This process is based on current guidelines for determining the frequency of assessments in [28], in which the frequency is reduced to 12 hourly in patients who are assessed to be at minimal risk of deteriorations, as indicated by a NEWS of zero. If it is

FIGURE 2.1: A wrist-worn wearable sensor: Nonin Medical's WristOx$_2$ ® 3150 measures heart rate (HR), arterial blood oxygen saturation (SpO_2), and the photoplethysmogram (PPG) continuously. Source: [5] (CC BY 4.0, DOI: 10.5281/zenodo.166546)

possible to safely reduce the frequency of manual assessments in low risk patients using this approach, then it could result in a considerable reduction in workload.

2.4 Current Evidence for the Clinical Benefit of Wearable Sensors

Wearable sensors must satisfy two key requirements to confer clinical benefit in the broad population of acutely-ill patients:

1. **Data coverage:** Wearable sensors must provide a sufficiently high real-time data coverage, where the data coverage is defined as the proportion of time for which high quality data are available. Achieving high data coverage in an ambulatory population is not straightforward for several reasons. These include: the presence of artifact due to movement or poor sensor contact; the discomfort of wearing sensors for prolonged periods; and, minimal capacity for both clinicians and patients to troubleshoot any problems, since clinicians already have busy workloads, and patients are often unfamiliar with wearable sensors.

2. **Multi-parameter monitoring:** Wearable sensors must acquire a sufficiently comprehensive set of signals and parameters to be able to detect the physiological changes indicative of deteriorations. Although a broad set of parameters are measured during intermittent observations (Table 2.1), it is difficult to acquire such a comprehensive set from wearable sensors. This is due to: the increased burden on the patient of wearing multiple

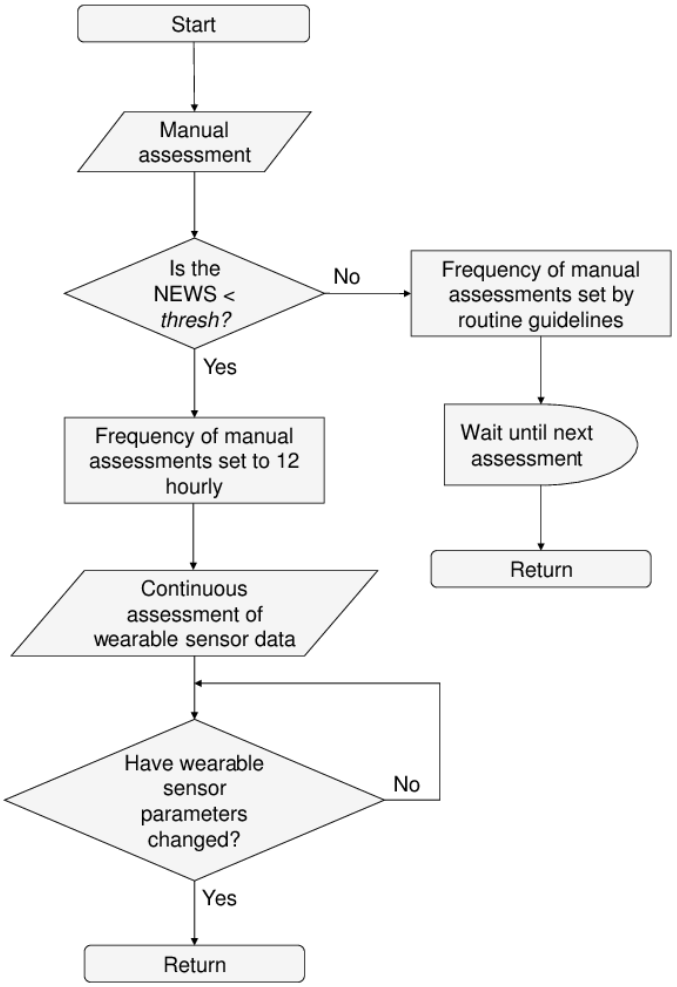

FIGURE 2.2: Using wearable sensors to reduce staff workload: This process could be used to reduce the workload associated with manual physiological assessments. An initial manual assessment is taken to determine whether a patient is at sufficiently low risk of deterioration, as indicated by a NEWS $<$ *thresh*. Low risk patients are continuously assessed for changes in a subset of physiological parameters using wearable sensors. If a change is detected then a manual assessment is triggered; otherwise, the delay between manual assessments is 12 hours, reducing workload.

transducers (such as both a pulse oximeter and a set of ECG leads); the increased power requirements and consequent shortening of battery life when using multiple transducers; and, the difficulty of acquiring some parameters accurately and unobtrusively, such as RR.

If wearable sensors satisfy these requirements, resulting in a high enough data capture rate for a sufficiently comprehensive set of physiological signals and parameters, then they may confer clinical benefit.

Surprisingly few studies have been conducted using multi-parametric wearable sensors in hospitals, as summarised in Table 2.4. Those studies that have been conducted are subject to the shortcomings of using short monitoring periods and small sample sizes, as illustrated in Figure

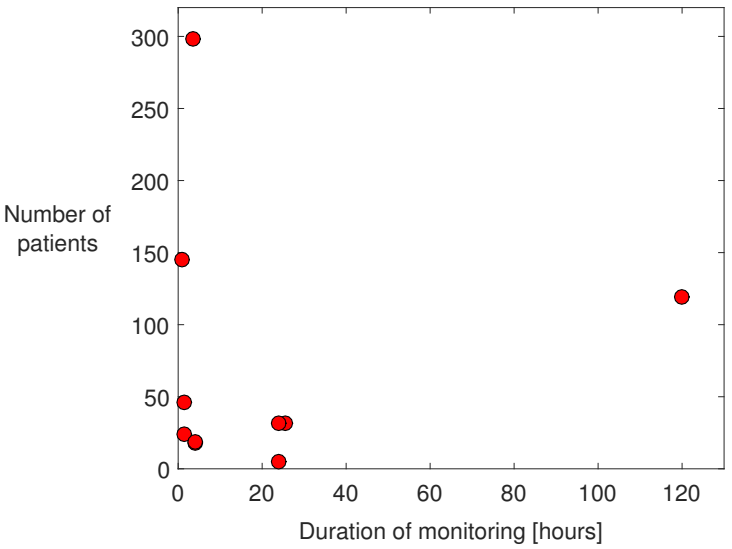

FIGURE 2.3: The shortcomings of studies of wearable sensors: Most studies of multi-parametric wearable sensors (represented by dots) have either monitored patients for short periods (typically 24 hours or less), or have enrolled few patients (typically less than 50). The studies shown are detailed in Table 2.4. It was assumed that the patients monitored *"until discharge"* in [85] were monitored for 120 hours (five days).

2.3. Evaluations of wearable sensors must be conducted over a similar time-scale to that for which they would be used in practice. Acutely-ill patients may stay in hospital for several days or weeks. However, all except one of these studies monitored patients for less than 26 hours on average, limiting the relevance of their findings with regard to patient acceptance of the sensors. Furthermore, evaluations must be adequately powered (*i.e.* have a sufficiently large sample size) to assess inter-subject variability in feasibility, and to identify clinical benefit given that only a small minority of patients are likely to deteriorate. However, only three studies enrolled more than 50 patients. This suggests that further research is required to assess the technical functionality and clinical benefit of wearable sensors.

Existing studies provide insight into the technical feasibility of acquiring data from wearable sensors in hospitals. Half of the studies identified in Table 2.4 reported packet reception ratios or data coverage rates, and are listed in Table 2.5. The packet reception ratios indicate the proportion of data transmitted from a wearable sensor which were successfully received. The reported packet reception ratios of over 95 % suggest that it is feasible to transmit and receive data from wearable sensors in the hospital setting. Indeed, in one study, Chipara *et al.* found that service intervals (periods of time for which transmissions were continuously received) lasted a median of 17.7 minutes, with 90 % of subsequent outage intervals (periods of time for which transmissions were not received) lasting 1.4 minutes or less [86]. Therefore, the challenges associated with transmitting wearable sensor data in a hospital, such as obstacles in wards, a

TABLE 2.4: Previous studies of multi-parametric wearable sensors in hospitals.

Study	Setting	Number of patients	Monitoring duration [hours]	Sensors	Monitored parameters
Curtis *et al.* [84]	-	145	0.8	ECG, pulse ox, geo-positioning	HR, SpO_2, location
Orphanidou *et al.* [66]	Acutely-ill inpatients	18	≤ 4.0	ECG, pulse ox, IP	HR, SpO_2, RR
Kisner *et al.* [85]	Post-surgical unit	119	until discharge (days)	pulse ox	HR, SpO_2
Chipara *et al.* [86, 87]	Acutely-ill inpatients	32	25.6	pulse ox	HR, SpO_2
Pollack *et al.* [88]	Emergency Dept	298	3.5	ECG	HR
Ko *et al.* [33]	Emergency Dept	46	1.5	pulse ox	HR, SpO_2
Ko *et al.* [89]	Pre-, peri-, and post-surgery	≤ 24	≥ 1.5	pulse ox	HR, SpO_2
Lopez *et al.* [90]	Cardiology unit	5	24.0	ECG, temperature, accelerometer	HR, angle of inclination, activity, temperature
Donnelly *et al.* [91]	Post-surgical unit, sleep laboratory, operating room	19	≈ 4.0	ECG, temperature, accelerometer, IP	HR, RR, temperature, activity
Donnelly *et al.* [91]	-	23	-	ECG, temperature, accelerometer, IP	HR, RR, temperature, activity
Bonnici *et al.* [75]	Acutely-ill inpatients	32	24	ECG, pulse ox	HR, SpO_2
Dor *et al.* [92]	-	-	-	pulse ox	HR, SpO_2

TABLE 2.5: Studies of wearable sensors which reported data acquisition rates: Studies reported packet reception ratios (the proportion of data packets transmitted from a wearable sensor which were successfully received) and data coverage rates (the proportion of time for which valid data were acquired). ** indicates that multiple devices were tested.*

Study	Sensor(s)	Packet reception ratio [%]	Data coverage rate [%]
Orphanidou *et al.* [66]	ECG	-	97
	pulse ox	-	88
Chipara *et al.* [86, 87]	pulse ox	99.7	≈ 80.9
Ko *et al.* [33]	pulse ox	95.4	-
Ko *et al.* [89]	pulse ox	98.3	-
Lopez *et al.* [90]	accelerometer ECG, thermometer	≈ 99	-
Bonnici *et al.* [75]	ECG	-	56.1, 38.8 *
	pulse ox	-	18.7, 56.1, 9.9 *

changing environment due to movements of people and equipment [33], and transmission within close proximity to the human body, do not appear to preclude the use of wearable sensors.

Evidence for the feasibility of using wearable sensors to acquire physiological measurements from hospital patients is less conclusive. Three studies reported the data coverage rate, the proportion of time for which valid data were acquired. Data can be rendered invalid by sensor disconnection, poor sensor connection or motion artifact. A wide range of data coverage rates have been reported for both ECG and pulse oximetry sensors, ranging from 19 to 97 % (see Table 2.5). This is likely to be indicative of the user acceptability of wearable sensors, since sensors may be removed by patients or staff if they are too burdensome. Whilst some studies have reported acceptability feedback from both patients and clinicians, none of these studies have monitored patients for over 24 hours [33, 88, 91]. It should be noted that the highest data coverage rates were achieved when monitoring patients for $\leq$ 4.0 hours [66]. In contrast, the lower data capture rates reported in [75, 86, 87] were achieved when monitoring patients for approximately 24 hours. This suggests that the acceptability of wearable sensors, and therefore the data capture rates, may be associated with the time for which patients have worn the sensors. Therefore, further research is required to determine whether the use of wearable sensors for prolonged monitoring is acceptable to patients.

Evidence for the clinical benefit of wearable sensors is sparse. The utility of wearable sensors

for arrhythmia detection has been reported [88, 91], although a high false alarm rate for critical arrhythmias has also been reported [84]. Only minimal evidence is currently available to support the hypothesis that wearable sensors can be used to detect deteriorations other than arrhythmias and result in changed patient outcomes. For instance, Kisner *et al.* reported that the use of a wearable pulse oximeter, coupled with a protocol for initiating supplementary oxygen therapy if the SpO_2 fell below 90 %, reduced the incidence of atrial fibrillation in patients recovering from a coronary artery bypass graft [85]. There are two reasons for a lack of evidence for the clinical benefit of wearable sensors. Firstly, very few studies have been conducted on the topic. Secondly, development of wearable sensors has focused largely on the hardware required to acquire streams of clinical data. However, simply acquiring data will not have an impact unless it is appropriately processed, providing decision support to clinicians. The next section will discuss how software for processing of wearable sensor data could be designed to realise the potential of wearable sensors.

2.5 Realising the Potential of Wearable Sensors

The key to realising the potential of wearable sensors for detecting deteriorations is found in the development of techniques to process and synthesise the continuous data to provide clinically useful and actionable outputs. This section provides the clinical background to the contributions of this thesis towards realising the potential of wearable sensors.

2.5.1 Development and evaluation of techniques for continuous, unobtrusive RR monitoring

RR has been found to be a highly sensitive marker of deteriorations [93]. For instance, cardiac arrests are preceded by abnormal RRs [18, 49, 94–96]; elevated RR can indicate respiratory dysfunction [97]; and, both elevated and reduced RRs ($\geq$ 20 and $<$ 6 breaths per minute, bpm) are predictors of mortality [22, 98, 99]. RR has been shown to be a particularly important indicator of deterioration. For instance, it has been observed to be the best discriminator of patient risk [31] and the best single variable predictor of cardiac arrest [49]. Furthermore, it has been found to be highly predictive of resuscitation events and death (odds ratio, OR (95% CI) = 9.1 (7.12 - 11.59) for RR $>$ 24 bpm) [23], and of in-hospital mortality (OR (95% CI) = 3.18 (2.89 - 3.50) for RR $>$ 22 bpm). Thus, there is much evidence that manual, intermittent

measurements of RR are highly sensitive markers of deteriorations, suggesting that RR has an important role to play in the detection of deteriorations using wearable sensors.

Current methods for taking individual and continuous measurements of RR are far from ideal, despite the clinical importance of RR. In current clinical practice individual measurements are usually taken by manually counting the chest movements associated with individual breaths over a period of approximately 30 - 60 s [100]. This practice is time-consuming, inaccurate [32], and poorly carried out [101–104]. Electronic methods for monitoring RR continuously are also susceptible to criticism, being obtrusive and inaccurate [32]. For instance, three methods are currently employed in wearable sensors to monitor RR: impedance pneumography (ImP), inductance plethysmography (InP), or an accelerometer. ImP is the least obtrusive of these methods since it consists of measurement of variations in thoracic impedance with respiration through injection of a high frequency voltage into the thorax at ECG electrodes [105]. However, ImP has been found to be imprecise, with 95% limits of agreement between reference RRs and ImP-derived RRs reported as ± 11.9 bpm and -9.9 - 7.5 bpm [32, 106]. These errors are large considering that a normal RR range is approximately 12 - 20 bpm. The use of InP is limited by the cumbersome nature of the chest bands which are required to hold a strain gauge against the thorax. Chest bands have been observed to be too uncomfortable for prolonged monitoring [66]. Until recently the use of accelerometers was confined to experimental devices [107], and their performance remains unclear. Therefore, there is an urgent need to develop a method for continuously and unobtrusively monitoring RR using wearable sensors.

A potential solution is to estimate RR from the ECG or PPG signals. Both signals are modulated by respiration as shown in Figure 2.4, providing opportunity to estimate RR from them. Furthermore, most wearable sensors monitor at least one of the ECG and PPG signals, since they are required to measure HR, and the PPG is required to measure SpO_2. Therefore, estimation of RR from the ECG or PPG would allow RR to be monitored without the need for any additional sensors, providing a valuable marker of deterioration. Over 100 algorithms have been proposed in the literature to estimate RR from the ECG and PPG [10]. However, it is not clear which of these algorithms provides the best performance, nor whether any of these algorithms perform sufficiently well for use in wearable sensors. This is because previous comparison studies have compared only a small number of algorithms [10] (see for example [108–113]). In addition they have often evaluated algorithms on data acquired from mechanically ventilated patients whose respiratory mechanics are different to those of ambulatory patients. Therefore, a

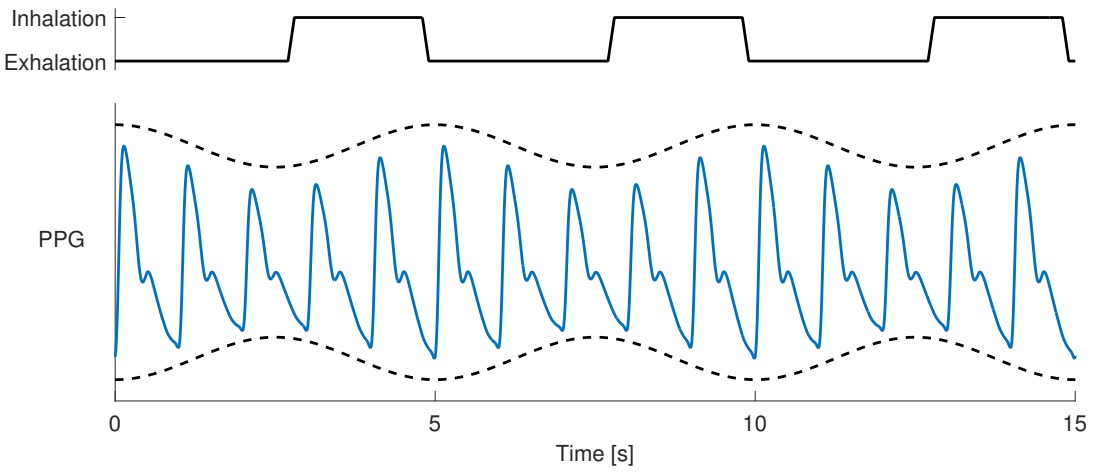

FIGURE 2.4: Modulation of the pulse oximetry (PPG) signal by respiration: Stroke volume varies during respiration due to changes in intrathoracic pressure and venous return. This causes amplitude modulation of the PPG signal. Other modulations of the PPG are also caused by respiration, as detailed in Section 4.2.

comprehensive evaluation of existing algorithms is required to identify the best algorithm, and determine whether its performance is sufficient for use in wearable sensors.

2.5.2 Identification of deteriorations from wearable sensor data

The second step in the processing of wearable sensor data is to generate alerts of deteriorating patients which can be used to trigger clinical actions. There has been much research into EWSs for identification of deteriorations from intermittent measurements. However, only minimal research has been conducted into techniques for detecting deteriorations from continuous wearable sensor data. Two techniques are in clinical use with wearable sensors. The first technique is to use an EWS to detect deteriorations with the continuous data provided by wearable sensors. This approach is now clinically available in the IntelliVue Guardian Solution (Philips Medizin-Systeme GmbH, Böblingen, Germany) [77], which can be customised to use any desired EWS algorithm [114]. Secondly, novelty detection algorithms are being used to assess the likelihood of deterioration using wearable sensor data. For instance, the Visensia^{TM} system, designed by Tarassenko *et al.* and commercialised by OBS Medical Ltd (Abigndon, UK), uses novelty detection methods to quantify the level of abnormality of a set of physiological parameters, indicating deteriorations [115]. These two techniques are now explained in further detail.

The use of an EWS to assess the likelihood of deterioration is a benchmark method for converting physiological data into clinically meaningful alerts. However, EWSs are designed for use with high quality data obtained from patients at rest during intermittent physiological assessments.

In contrast, wearable sensor data present additional challenges due to the increased likelihood of low quality data, and the possibility of data being acquired during physical activity. Both scenarios result in the physiological inputs to an EWS not being representative of a patient's true physiology when at rest. Furthermore, EWSs typically require inputs which are not provided by wearable sensors, meaning that additional inputs from the previous set of intermittent measurements are required, which may no longer be accurate. Therefore, the straightforward use of an existing EWS to synthesise wearable sensor data cannot be presumed to be an optimal solution. Modifications to existing EWSs may be needed, such as a requirement that elevated EWSs be sustained for a certain period before raising an alert to avoid false alerts due to transient, unrepresentative input data. However, an advantage of the use of EWSs is that they are already familiar to hospital clinicians due to their widespread use.

The second technique of using novelty detection algorithms to identify deterioration is more easily adapted for use with wearable sensor data. The Visensia^{TM} system consists of a multivariate probability density function (PDF) learnt from a historical dataset. An alert is triggered when the probability associated with a set of measurements is below an empirically-optimised threshold. The system was originally trained using continuous monitoring data [115]. Therefore, it is likely that the model is more robust to the potentially unrepresentative data provided by wearable sensors. In addition, the model was generated using statistical methods, unlike EWSs which usually contain an element of expert opinion [116]. Therefore, the system could easily be re-trained for use with wearable sensor data. Furthermore, the PDF can be generated from any desired set of numerical inputs. Consequently, this approach is suitable for use with wearable sensors which only provide a subset of the routinely measured physiological parameters. It is also suitable for patient-population specific monitoring, since a new PDF could be generated for each patient population (although this is not currently implemented). In addition, this approach may facilitate analysis of physiological variability, since the variability in each parameter, and the variability in their relationships, is captured by the PDF.

Techniques to identify deteriorations from wearable sensor data should be designed with the following three requirements in mind. Firstly, they must provide sensitive alerts of deteriorations. Secondly, the alerts must provide sufficiently earlier warning of deteriorations than current practice to be clinically advantageous. Thirdly, the positive predictive value (PPV) of alerts (the proportion of alerts which are true) must be sufficiently high to avoid alarm fatigue. The problem of alarm fatigue arises where staff have to respond to an overly high false alert rate. Previous studies of wearable sensors have identified extremely high false alert rates [64],

which has potential to desensitise staff to alarms, cause mistrust of the equipment, and result in a lack of response to true alarms [117–119]. Further research is required to determine whether either of the approaches described meets these three requirements, or whether novel techniques are required.

2.6 Final Remarks

Acutely-ill hospitalised patients are at risk of deterioration, defined as an acute worsening in health, resulting in transition to a physiological state with an increased likelihood of a clinical adverse event. Current practice for identification of deteriorations is based on intermittent physiological assessments, which are performed by hand at 4 - 6 hourly intervals. Recent developments in wearable sensors mean that they may now have a role in improving the detection of deteriorations, and reducing the workload associated with manual physiological assessments. The continuous data provided by wearable sensors may facilitate earlier identification of changes in physiology, personalised alerting, and analyses of physiological variability. However, very few studies have investigated the use of multi-parametric wearable sensors in hospitals. Most previous studies have been limited by low sample sizes or a short monitoring duration. Therefore, evidence for the feasibility of using wearable sensors to monitor hospitalised patients for prolonged periods is lacking, and evidence for the clinical benefit of wearable sensors is even more sparse. Consequently, further research is required before wearable sensors can be used for detection of deteriorations in hospital patients.

Two steps towards realising the potential of wearable sensors for detecting deteriorations have been identified. Firstly, a technique is required for continuous and unobtrusive RR monitoring using wearable sensors. This is because RR is a sensitive marker of deteriorations, yet current methods for monitoring RR using wearable sensors are highly obtrusive or inaccurate. Secondly, a technique is required to identify deteriorations from multivariate, noisy, wearable sensor data. Two existing techniques are now in clinical use: EWSs and novelty detection algorithms. However, it is not clear whether they perform well enough to meet the three requirements of a technique to identify deteriorations: sensitivity, advanced warning, and a high PPV. The remainder of this thesis presents investigations into continuous RR monitoring techniques for use with wearable sensors. In Chapter 8 a RR algorithm is used with algorithms to detect deteriorations to assess the utility of wearable sensors for detection of deteriorations.

Chapter 3

Assembling a Physiological Dataset

This chapter describes the processes undertaken to assemble a comprehensive physiological dataset acquired from 226 adult patients. Firstly, the clinical trial in which the data were collected is summarised. This is followed by descriptions of the processes for data preparation (preparation specific to this particular trial) and quality assessment (identification and elimination of unreliable continuous data). A detailed description of the dataset is then provided. Finally, the utility of the dataset for assessing the utility of wearable sensors is evaluated.

3.1 The Feasibility and Efficacy of Continuous In-Hospital Patient Monitoring (LISTEN) Trial

The Feasibility and Efficacy of Continuous In-Hospital Patient Monitoring (LISTEN) Trial was an observational study in which acutely-ill patients were monitored continuously using wearable sensors. Its primary aim was to determine whether wearable sensors could be used to provide earlier warning of deteriorations than standard practice. Secondary aims included assessments of the feasibility and clinical utility of using wearable sensors to monitor hospital patients continuously, and to develop novel early warning algorithms for use with wearable sensors. A highly comprehensive physiological dataset was acquired during the trial, including electrocardiogram (ECG) and photoplethysmogram (PPG) data from wearable sensors, making it particularly suitable for this work. The trial was approved by the Bloomsbury Research Ethics Committee (reference 12/LO/0526), and was carried out at Guy's and St Thomas' NHS Foundation Trust, London, UK (National Clinical Trial no. 01549717) between November 2012

and January 2014. Continuous physiological data were acquired throughout patients' post-surgical hospital stays, as well as routine health record data. Furthermore, labels of adverse events (defined in Table 8.1) were added retrospectively. Consequently, the dataset is a highly comprehensive record of the physiology of a set of acutely-ill hospital patients, both in terms of the frequency and breadth of measurements.

The patient population chosen for the trial was that of adult patients recovering in hospital after major cardiac surgery, which had both advantages and disadvantages. The advantages stemmed from the relative ease of acquiring continuous physiological data from this population. Firstly, it was possible to equip all of the wards on which these patients would typically stay with continuous monitoring equipment, since the patients typically followed a narrow and predictable clinical pathway. Secondly, all of the study wards already used continuous monitoring equipment when clinically indicated. Therefore, ward staff were familiar with the equipment, making it easier to use continuous monitoring equipment with the study patients. However, there were two key disadvantages. Firstly, the incidence of deteriorations in this population is low (in comparison to high-risk surgical cohorts). Approximately 10-20% of patients recovering from cardiac surgery suffer from a deterioration during their hospital stay [38, 39], and their hospital mortality rate is approximately 3% [120]. In contrast, approximately 50% of patients recovering from major gastrointestinal surgery suffer from a deterioration [37]. Secondly, clinical staff could not be blinded to the continuous physiological data, since the monitoring equipment was already in routine clinical use. Therefore, it may have helped staff to identify deteriorations earlier, prompting clinical interventions which would otherwise not have been carried out.

Patients recovering from major cardiac surgery are at risk of a wide range of adverse events. The adverse events which can be expected after coronary artery bypass graft (CABG) surgery, a common type of cardiac surgery [120], have been reported in [40]. Several organ systems are affected by these events. For instance, the heart is affected by congestive heart failure and other cardiac events; the cardiovascular system is affected by bleeding and hypotension; the respiratory system is affected by pneumonia, respiratory compromise, and pneumothorax; and the kidneys are affected by renal dysfunction. Consequently, the physiological manifestations of the deteriorations preceding these events can be expected to differ greatly with different types of event. For instance, cardiac deteriorations are commonly manifested as abnormalities on the ECG. In contrast, respiratory deteriorations primarily impact vital signs such as RR and SpO_2, whereas cardiovascular deteriorations primarily impact HR and BP. Since each type of deterioration may affect a different organ system, and manifest differently, it may be necessary

TABLE 3.1: The typical clinical pathway of patients participating in the LISTEN trial whilst recovering from cardiac surgery.

Care Level	Description	Duration [days], med (quartiles)	Monitoring Equipment	Continuous Signals	Parameters
1-3	Hospital Care	7.0 (5.3 - 11.1)	n/a	n/a	n/a
1	Critical Care	0.9 (0.7 - 1.3)	Static bedside monitors	ECG, PPG, ImP, CVP, ABP	HR, SpO_2, RR, CVP, SBP, DBP, MAP
2	Critical Care	1.3 (1.0 - 2.3)	Static bedside monitors	ECG, PPG, ImP, CVP	HR, SpO_2, RR, CVP
3	Ambulatory	4.2 (3.1 - 8.1)	Wearable telemetry monitors	ECG, PPG	HR, SpO_2

to design algorithms to detect individual types of deteriorations, or groups of deteriorations which manifest similarly.

Patients in this trial typically stayed in two distinct clinical settings with differing capabilities for continuous physiological monitoring: critical care and the ambulatory ward. Patients were initially treated in a level 1 critical care ward following surgery, known as an intensive care ward. They were then cared for in a level 2 critical care ward, known as a high dependency ward. Finally, they stayed in a level 3 ambulatory ward until hospital discharge. The first two settings, the level 1 and 2 critical care wards, used static bedside monitors (IntelliVue MP70 monitor, Philips Medical Systems, Andover, MA, USA) for continuous physiological monitoring. The physiological signals monitored were lead II ECG, finger PPG, impedance pneumography (ImP), arterial blood pressure (ABP) signals (although ABP was rarely monitored in the level 2 ward), central venous pressure (CVP), and occasional additional ECG leads. In contrast, in the level 3 ambulatory ward static monitors were no longer used, and instead wearable sensors (IntelliVue TRx M4841 telemetry monitors, Philips Medical Systems) were used, which only provided continuous monitoring of ECG and PPG signals. Further details of the signals and parameters are provided in Table 3.1. As a result of the reduced monitoring capabilities on the ambulatory ward, the only continuous physiological signals available throughout patients' recoveries were ECG and PPG, and the only continuous parameters were HR and SpO_2.

Continuous monitoring data, which consisted of physiological signals, parameters and monitoring alarms, were acquired from the bedside monitors and the wearable sensors using BedMaster data acquisition software (v.4.1.12, Excel Medical Electronics, Jupiter, FL, USA). The software associated data with the monitor from which they were recorded, rather than with a particular patient. Therefore, the times at which patients were being monitored by each monitor were manually recorded in the electronic Case Report Form (eCRF) to allow the data associated with each patient to be identified. The data were recorded in a proprietary format, and converted to Extensible Markup Language (XML) format using the BedMaster software. They were then converted to Matlab ® format using bespoke software.

The acquisition of intermittent physiological measurements also varied between clinical settings. In the level 1 and 2 critical care wards observations were routinely recorded in an Electronic Health Record (EHR), and could be exported to a text file for subsequent processing. However, in the level 3 ambulatory ward observations were routinely recorded on paper notes. In an attempt to capture these electronically we introduced new vital signs monitors (CVSM 6000, Welch Allyn, UK) to the ward for the trial. This allowed the user to optionally send measurements directly from the monitor to an EHR, in addition to the routine use of paper notes. However, this did not provide a comprehensive record of the intermittent measurements on its own, since many measurements were not sent from the monitor to the EHR. Therefore, measurements from the ambulatory ward were dual-transcribed from both the electronically transmitted data and the paper notes, to obtain as complete a dataset as possible. The intermittent measurements will serve two purposes in this thesis. Firstly, they will be used to calculate benchmark early warning scores (EWSs), which will be used as comparators when evaluating novel algorithms for detection of deteriorations. Secondly, they will be used as inputs to novel algorithms, supplementing the continuous data with additional vital signs which are not monitored continuously, such as temperature.

Three further types of data were recorded. Firstly, fixed variables, *i.e.* those which do not change during a patient's hospital stay, were manually recorded in the eCRF. These included demographic data and operation details. They may have utility as inputs to early warning algorithms to stratify patients into groups [121]. Secondly, laboratory test results were automatically recorded in an EHR as part of routine care. They were retrospectively exported as text files. They may also be useful inputs to early warning algorithms since algorithms have recently been developed which incorporate laboratory values in addition to traditional vital sign

inputs resulting in improved performance [122]. Thirdly, adverse events were identified retrospectively from each patient's notes by an expert and recorded in the eCRF. Best estimates of the times of events were recorded, as well as the times at which they were identified by clinical staff. These labels are essential for this work since they provide the response variable against which early warning algorithms can be trained and assessed.

For further details of the LISTEN Trial the reader is referred to [35, 123].

3.2 Data Preparation

The first stage of dataset curation was to perform the data preparation tasks which were specific to this particular dataset. The aim was to convert the raw data files into a *Prepared Dataset* on which analyses can be performed. The following tasks were performed:

- **Fixed variables:** Age, gender, ethnicity, operation type and operation duration were extracted from the eCRF. Ethnicities were grouped as: Asian, Black, White, or Other. Operation types were grouped as: bypass, valve, bypass and valve, or other.
- **Events:** The following events were extracted from the eCRF: the wards which a patient stayed on; the times at which patients were admitted to and discharged from each ward; the times at which patients were provided with a wearable sensor; and the type and times of any withdrawals from the trial. In addition, the times at which patients stayed in critical care, an ambulatory ward, and the study ward (the ambulatory ward on which patients were provided with wearable sensors) were extracted.
- **End-points:** The types and times of adverse events were extracted from the eCRF, as well as the following additional end-points: re-admission to critical care; hospital discharge (regardless of survival status); and discharge from hospital alive.
- **Laboratory test results:** The majority of laboratory tests gave a result in either a numerical or a categorical value, whereas a small minority of results were in free-text format. Those tests which had numerical or categorical results were extracted from the raw text files. Both the times of tests and the results were extracted. Tests with free-text results were not exported.

- **Intermittent physiological measurements:** Intermittent physiological measurements were extracted from EHR exports (corresponding to critical care) and dual-transcribed spreadsheet files (ambulatory ward). Most sets of measurements contained the five vital signs (HR, SpO_2, SBP, temp and RR) and O_2. In addition, measurement sets collected in critical care contained the heart rhythm and ventilator settings.

- **Continuous monitoring data curation:** Any continuous monitoring data with either implausible timestamps (*e.g.* 01-Jan-1970), or physiologically implausible values (*e.g.* values $> 1 \times 10^9$), were identified as erroneous and discarded.

- **Continuous monitoring alarms:** Continuous monitoring alarms were categorised according to: (i) whether they were physiological; (ii) the input signal; (iii) type (such as 'threshold limit breached'); and (iv) level of severity (as provided by the monitoring equipment).

- **Timestamps:** All timestamps were adjusted to be relative to the time at which the patient left the operating theatre. Times were measured in seconds (except for the times of physiological signals, which were measured in milliseconds).

- **Formatting:** All data were imported into Matlab ® format (*.mat*). This provided efficient data storage, and ensured that all the data were accessible for analysis via a single software platform.

- **Correction of continuous monitoring data timestamps:** The timestamps of continuous monitoring data were sometimes incorrect, perhaps due to latencies in data transmission prior to timestamping. This resulted in sections of data being shifted in time, resulting in data gaps and simultaneous data, as illustrated in Figure 3.1(a). A bespoke algorithm was designed and implemented to correct the timestamps. This resulted in correction of the timestamps of many of the shifted sections, as demonstrated in Figure 3.1(b). The results of using the algorithm are summarised in Table 3.2. The key advantage of using the algorithm is that it increases the number of segments which contain continuous data, allowing additional parameters to be derived from the waveforms. Further details of the algorithm are available in [35].

The *Prepared Dataset* is designed to be an accessible representation of the raw data from the LISTEN trial, following essential data curation processes. It is useful for a small set of analyses which require all the data, regardless of its quality. For instance, the primary aim of the

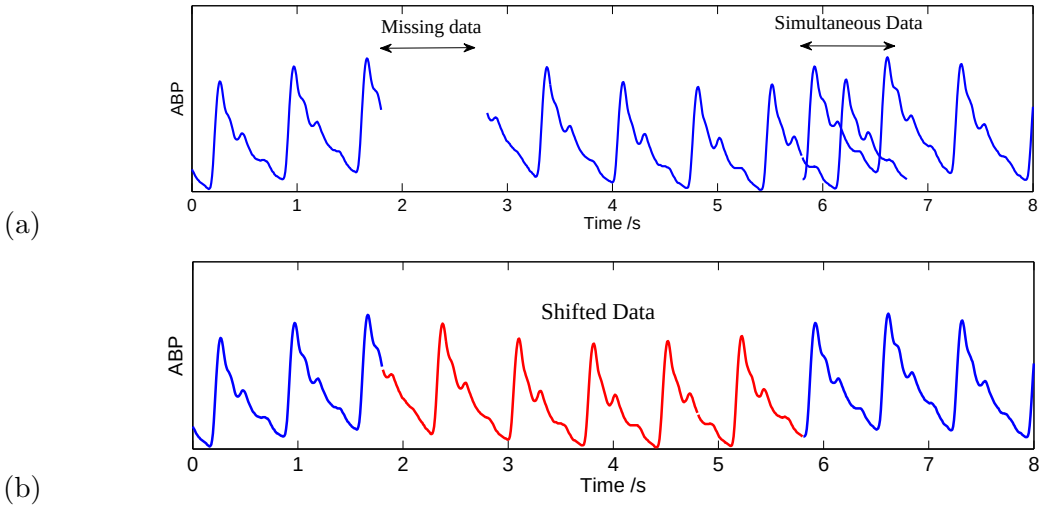

FIGURE 3.1: Correction of continuous monitoring data timestamps: Errors in timestamps resulted in simultaneous data at some times, and missing data at others, as shown in (a). A bespoke algorithm was implemented to correct for erroneous timestamps, as shown in (b). Figure and caption adapted from [35].

TABLE 3.2: Performance of the timestamp correction algorithm: the proportion of data points, and continuous data segments, available for analysis are shown before and after using the timestamp correction algorithm. The key benefit of the algorithm is the increased percentage of continuous data segments available for signal processing. Assessed using the first 24 hours of heart rate data acquired from each subject after admission to critical care following surgery.

	Proportion available [%], med (lower - upper quartiles)	
	Before algorithm	After algorithm
Data points	94.2 (91.3 - 98.0)	95.3 (92.4 - 99.0)
10 s segments	90.5 (88.7 - 92.0)	94.6 (91.4 - 98.0)
30 s segments	85.9 (83.8 - 88.6)	93.2 (89.6 - 96.1)
60 s segments	82.0 (77.8 - 85.0)	91.2 (87.2 - 94.0)

LISTEN trial was to assess the feasibility of using wearable sensors for continuous monitoring. The *Prepared Dataset* should be used for analyses such as this. However, most analyses are best conducted using only those data which are of high quality. As such, they should be conducted on the *Processed Dataset*. This is an enhanced version of the *Prepared Dataset* in which data deemed to be of low quality have been discarded. The processing steps performed to obtain the *Processed Dataset* from the *Prepared Dataset* are now described.

3.3 Quality Assessment

The second stage of dataset curation consisted of eliminating unreliable continuous monitoring data from the *Prepared Dataset*. Unlike intermittent physiological measurements, continuous monitoring data are recorded without verification by a clinician. Therefore, they are susceptible to inaccuracies due to factors such as movement artifact and poor sensor contact, making quality assessment particularly important. Several methods have been proposed in the literature to identify artifactual data [114]. In this section the methods used to identify and eliminate low quality continuous data are described. The use of quality assessment methods ensured that only high quality continuous data were retained in the *Processed Dataset*. All data other than continuous monitoring data were simply carried forward from the *Prepared Dataset* to the *Processed Dataset*.

Quality assessment was performed by calculating signal quality indices (SQIs) for ECG, PPG, ABP and ImP signals. SQIs are metrics derived from a physiological signal which indicate its quality. In this work SQIs were calculated for the ECG and PPG after segmentation into windows of 10 s duration [124]; for the ABP using 20 s windows [125]; and for ImP using 32 s windows (as described in Section 4.6.2). A range of measurements indicative of signal quality were derived from each signal and each window. The quality of each window was then determined through comparison of the measurements to empirical threshold values, indicative of the boundaries between low and high quality. Low quality windows of signal were discarded, as were any parameters derived from these windows (such as SBP values derived from a window of ABP signal). The quality assessment process is helpful for avoiding false alarms in early warning algorithms, which are often caused by parameters being inaccurately derived from artifactual signal data [126]. An example is provided in Figure 3.2, where false bradycardia and desaturation alerts were triggered by low quality data, which were eliminated through quality assessment. The signal processing steps involved in calculation of SQIs are now described.

The first step towards calculation of SQIs was to detect heart beats in the ECG, PPG and ABP signals, and breaths in the ImP signal. Beat detection was performed on the ECG using the QRS detector proposed by Pan, Tompkins and Hamilton [72, 73]. Two implementations of this QRS detector were used. Firstly, the implementation that was used in the original SQI algorithms was used to facilitate SQI calculation [124]. However, since this implementation has not been rigorously tested and is not publicly available, a second implementation, `rpeakdetect.m`, was

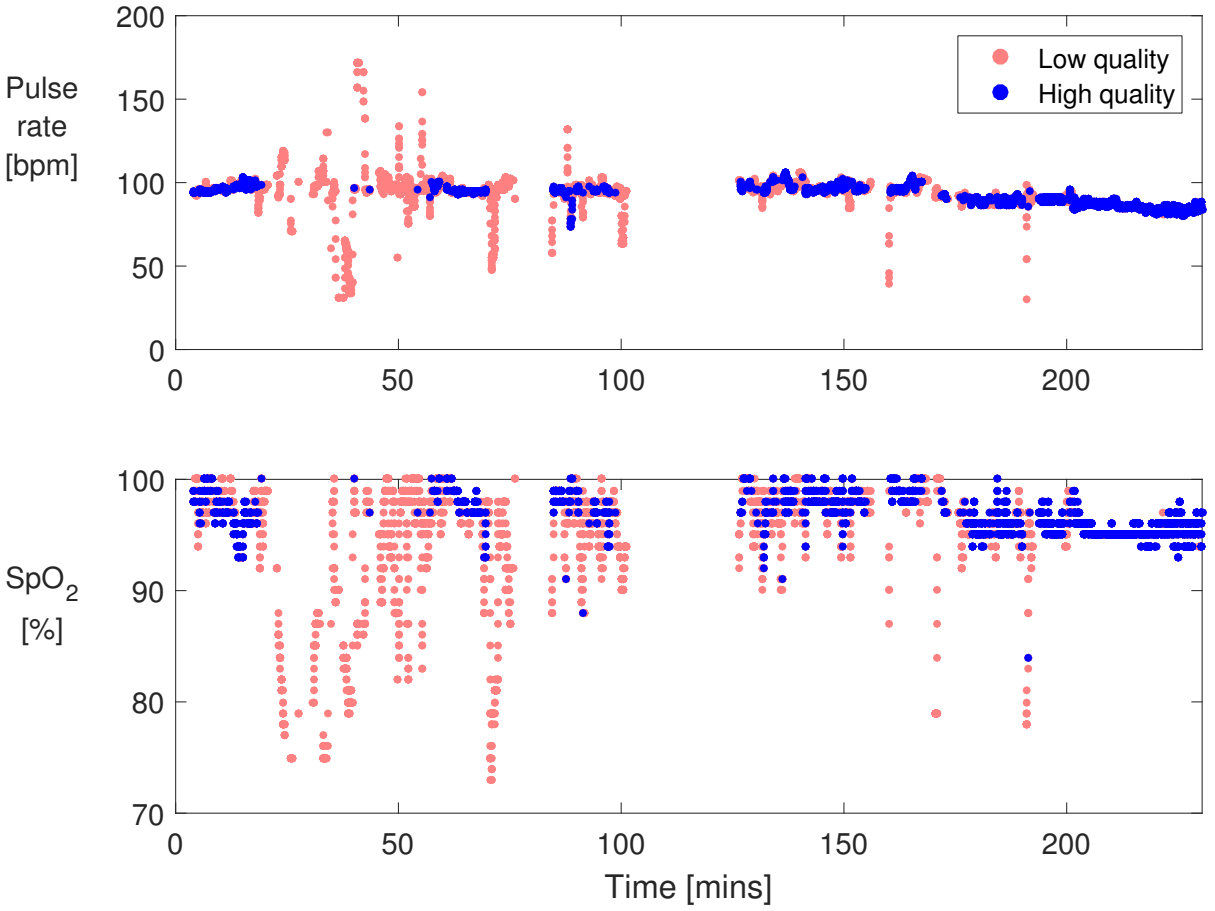

FIGURE 3.2: Quality assessment of PPG-derived numerics. The pulse rate (PR) and arterial blood oxygenation (SpO_2) numerics provided by the clinical monitor are shown. It appears that the patient desaturated at approximately 35 and 70 mins, and that the pulse rate dropped at 40 mins. However, quality assessment of the PPG waveform from which these numerics were derived revealed that the numerics coloured in pink were of low quality, indicating that these derangements were not physiological, but artifacts of low quality data.

used for all other purposes besides SQI calculation [127]. Similarly, PPG SQI calculation was performed using the three-point peak detector used in the original SQI algorithms [124]. For all other purposes the Incremental Merge Segmentation algorithm proposed in [74] was used for beat detection. This particular implementation was written by M. Pimentel [128]. Beat detection in the ABP signal was performed using the pulse onset detector described in [129], which was implemented in the code accompanying [125]. A modification of this code was used for both SQI calculation and other purposes. Breath detection in the ImP signal was performed using the *count-orig* method, which was originally proposed for use with respiratory signals derived from the ECG [109]. This consists of: (i) detrending the signal; (ii) detecting peaks and troughs; (iii) defining a threshold as 0.2 times the 75th percentile of peak values; (iv) ignoring peaks with an amplitude below this threshold; and, (v) identifying valid breaths as consecutive peaks separated by only one trough with an amplitude less than zero.

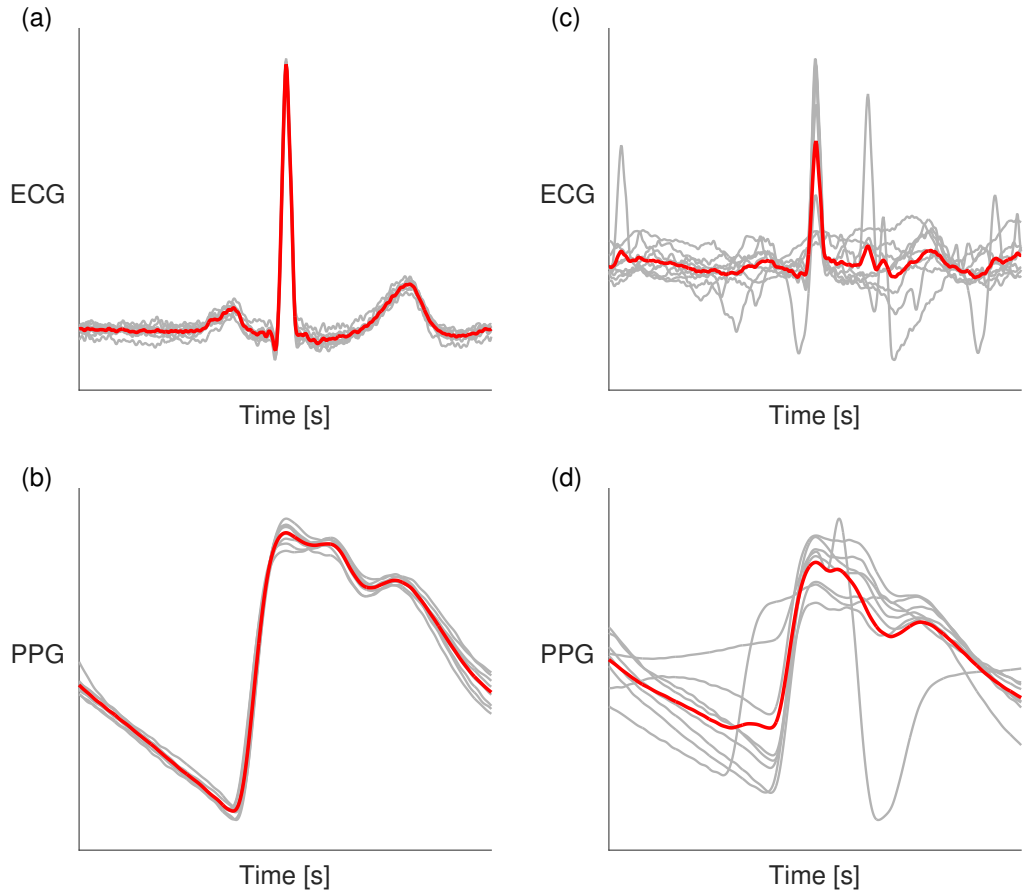

FIGURE 3.3: Template-matching for signal quality assessment: ECG and PPG signals were segmented into windows, and the correlation between individual beats (thin grey lines) in a window and the window's average beat template (thick red line) was calculated. If the correlation was below an empirically determined threshold then the segment was deemed to be of low quality, as described in [124]. (a) and (b) show high quality windows, whereas (c) and (d) show low quality windows. Figure and caption adapted from [10].

The second data processing step was to perform signal quality analysis on the continuous signals. Signal quality analysis of the ECG and PPG signals was performed using the template-matching algorithms proposed in [124]. The algorithms consist of two parts. Firstly, the HR and the beat-to-beat intervals are derived from the heart beat annotations. These are compared to a series of thresholds to determine whether the timings of the annotated heart beats are physiologically plausible. Secondly, a template beat is calculated as the average of each individual beat. The correlation between each individual beat and this template is calculated, and compared to an empirical threshold indicating the minimum acceptable correlation. If a window fails any of these tests then it is deemed to be of low quality. The template-matching process is demonstrated in Figure 3.3. These algorithms were chosen since they were designed for use with ambulatory data.

Signal quality analysis of the ABP signal was performed using the algorithm proposed in [125]. This consisted of deriving a range of metrics for each beat, and assessing whether they are physiologically plausible. The metrics included the systolic, diastolic and mean arterial blood pressures, and the beat duration. If any metric was outside of a pre-specified range of physiologically plausible values then that beat was deemed to be of low quality. If a window contained any low quality beats, then it was deemed to be of low quality. This algorithm is based on the assumption that physiologically plausible ABP signals can be described by a fixed set of metrics. Unlike the ECG and PPG SQIs, it does not adapt to an individual patient's signal morphology.

The remaining signal for which a SQI was required was the ImP (respiratory) signal. Signal quality assessment techniques for use with respiratory signals are far less established than those for use with cardiac signals. Therefore, a novel signal quality index (SQI) for the ImP signal was developed and assessed, as described in Section 4.6.2. This SQI completed the set of SQIs required to assess the quality of all of the signals from which routinely used vital signs are derived.

3.4 Dataset Description

A total of 229 subjects were enrolled into the LISTEN Trial. Three subjects subsequently withdrew consent so were excluded from any analyses. The demographic characteristics of the 226 remaining subjects are summarised in Table 3.3. A total of 201 subjects (89%) remained in wards where physiological data could be acquired continuously throughout their stay. Furthermore, 200 subjects (89%) were provided with a wearable sensor for 3.7 (2.6 - 5.2) days. The physiological dataset acquired from these subjects is presented as two distinct datasets. The *Prepared Dataset* is the dataset after completion of the data preparation described in Section 3.2, but before any of the data processing described in Section 3.3. The *Processed Dataset* is the dataset after completion of the additional processing described in Section 3.3. These datasets are described in this section to facilitate evaluation of the dataset for future research. Firstly, an evaluation of the utility of the dataset for this thesis is presented in Section 3.5. Secondly, the description in this section will allow future researchers to evaluate the utility of the dataset for their own research.

TABLE 3.3: Demographic characteristics of the 226 LISTEN subjects who underwent surgery

Characteristic	
Age [yrs], median (lower - upper quartiles)	67.5 (60.0 - 76.0)
Gender [male], no. (%)	166 (73.5)
Ethnicity, no. (%)	
White	210 (92.9)
Asian	6 (2.7)
Black	7 (3.1)
Other	1 (0.4)
Operation Type, no. (%)	
Bypass	71 (31.4)
Valve	111 (49.1)
Bypass and Valve	36 (15.9)
Operation Duration [hrs], median (lower - upper quartiles)	3.4 (2.8 - 4.2)

The LISTEN data is presented as the *Prepared Dataset* and the *Processed Dataset* since the two datasets are suitable for different types of research. The *Prepared Dataset* is designed to contain all of the recorded data, with the exception of nonsensical data, regardless of its quality. Consequently, it has the advantage of containing additional continuous signals which were not quality assessed, such as central venous pressure (CVP) and additional ECG leads, and their accompanying parameters. However, it has the disadvantage that the continuous physiological data have not been quality assessed. This dataset is designed for either: (i) the experienced analyst who wishes to process the data using their own methods, or (ii) research which requires 'real-world' data, as opposed to solely high quality data. The *Processed Dataset* is designed to be used for analyses which require only high quality data. It requires minimal pre-processing, since any low quality continuous physiological data have already been eliminated. It is designed for use by novices, and is particularly suitable for use with supervised machine learning algorithms since it is structured to facilitate extraction of predictor and response variables. A selection of end-points are provided, giving a choice of response variables. Figure 3.4 shows an example of the data available for a single patient in the *Processed Dataset*. All data other than continuous physiological signals and numerics are identical in the two datasets.

The data coverage rates for the continuous data were limited by the settings in which the data

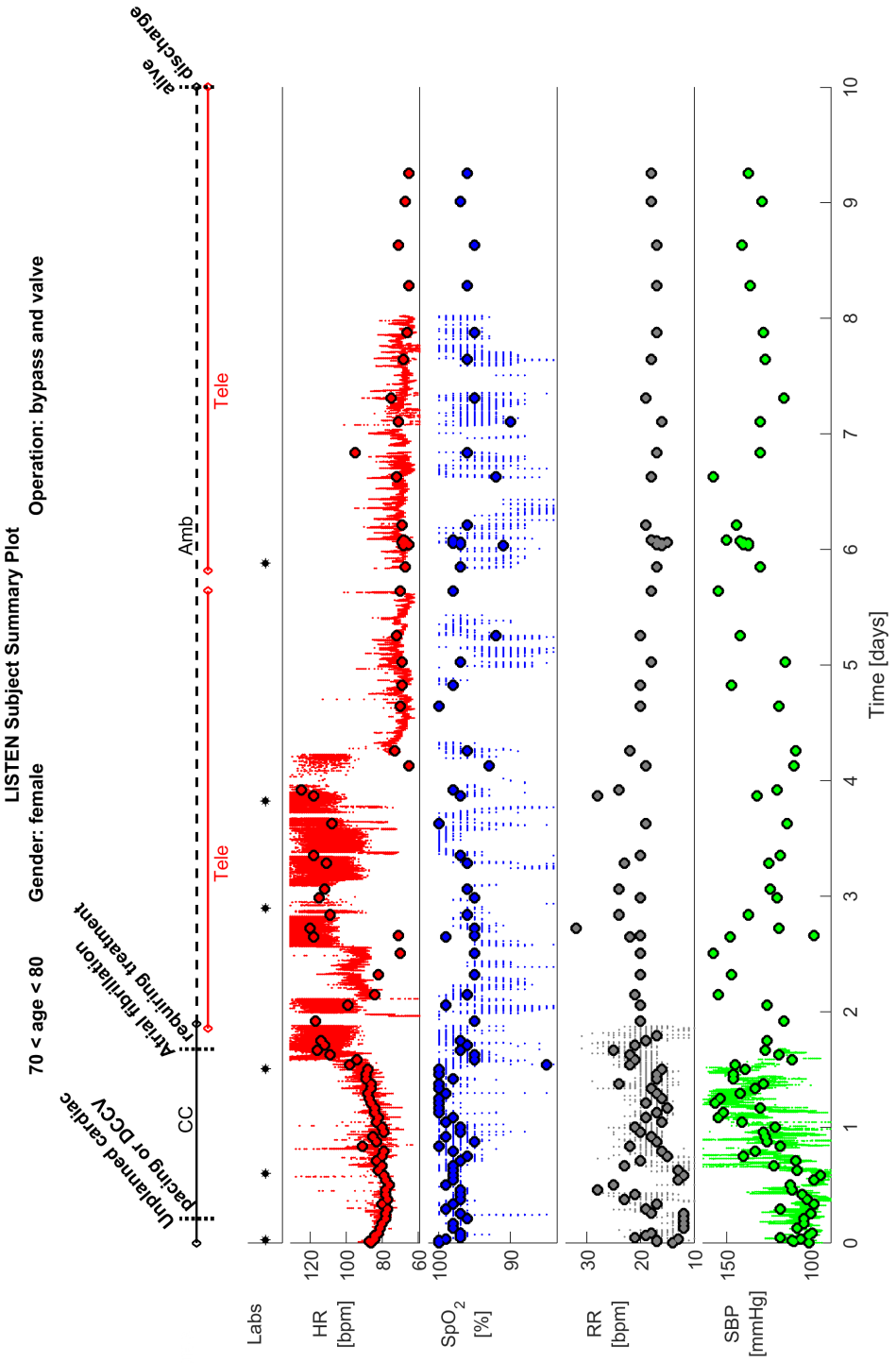

FIGURE 3.4: A summary of a LISTEN subject's processed data: Fixed variables (such as demographics) are shown at the top, followed by labels of adverse events. The times at which the patient was in critical care (CC), on the ambulatory ward (Amb), and being monitored by a wearable sensor (Tele) are then shown. This is followed by the times of laboratory tests (Labs), and then each of four vital signs: heart rate (HR), arterial blood oxygenation (SpO_2), respiratory rate (RR) and systolic blood pressure (SBP). Continuous data is shown in small dots, whereas intermittent measurements are shown as larger dots with a black outline.

TABLE 3.4: Data coverage rates of continuous data in the *Processed Dataset*. Definitions: *prop*: Proportion of time for which data were acquired [%], median (lower - upper quartiles); *lag*: Time until next measurement [median hrs], median (lower - upper quartiles).

Variable	Entire Dataset		Critical Care		Ambulatory	
	prop	lag	prop	lag	prop	lag
ECG	56.6 (39.0 - 70.6)	0.0 (0.0 - 1.9)	78.3 (68.7 - 82.6)	0.0 (0.0 - 0.0)	45.7 (15.7 - 67.1)	0.0 (0.0 - 8.1)
PPG	27.9 (17.6 - 37.7)	0.8 (0.0 - 51.5)	54.4 (42.8 - 64.3)	0.0 (0.0 - 0.0)	9.6 (1.1 - 26.2)	21.0 (1.1 - ∞)
ImP	5.3 (3.0 - 8.0)	∞ (0.6 - ∞)	15.6 (9.9 - 21.7)	0.2 (0.0 - 0.7)	n/a	n/a
ABP	6.8 (4.1 - 11.7)	∞ (∞ - ∞)	21.2 (11.5 - 30.6)	∞ (0.0 - ∞)	n/a	n/a
HR	56.4 (39.1 - 70.3)	0.0 (0.0 - 1.9)	78.5 (68.7 - 82.7)	0.0 (0.0 - 0.0)	45.4 (15.8 - 67.1)	0.0 (0.0 - 8.1)
PR	27.6 (17.4 - 37.7)	0.9 (0.0 - 164.6)	54.7 (43.1 - 64.2)	0.0 (0.0 - 0.0)	7.9 (0.9 - 25.8)	31.0 (1.6 - ∞)
SpO_2	27.6 (17.3 - 37.7)	0.9 (0.0 - 70.3)	54.5 (42.8 - 64.2)	0.0 (0.0 - 0.0)	8.4 (0.9 - 25.8)	27.3 (1.5 - ∞)
SBP	6.9 (4.2 - 11.9)	∞ (∞ - ∞)	21.4 (11.6 - 30.7)	45.1 (0.0 - ∞)	n/a	n/a
DBP	6.8 (4.2 - 11.9)	∞ (∞ - ∞)	21.4 (11.6 - 30.7)	45.1 (0.0 - ∞)	n/a	n/a
MAP	6.8 (4.2 - 11.9)	∞ (∞ - ∞)	21.4 (11.6 - 30.7)	45.1 (0.0 - ∞)	n/a	n/a
PP	6.8 (4.2 - 11.9)	∞ (∞ - ∞)	21.4 (11.6 - 30.7)	45.1 (0.0 - ∞)	n/a	n/a

were recorded. Continuous data could only be recorded from subjects whilst they were staying on *study wards*, *i.e.* those equipped with data acquisition equipment. Out of the 226 subjects, 25 (11.1%) stayed on a non-study ward at least once during their hospital stay. These subjects stayed on non-study wards for a median (lower - upper quartiles) of 43.2 % (26.5 - 59.7) of their hospital stay. In addition, only a subset of the continuous variables available in critical care could be acquired in the ambulatory setting as described in Section 3.1. The data coverage rates of the continuous data in the *Processed Dataset* are summarised in Table 3.4. These are shown firstly for the entire dataset, and then broken down according to critical care and ambulatory settings.

Several aspects of the continuous data coverage rates are useful for informing the design of algorithms for use with wearable sensors. Firstly, only the ECG and PPG signals were available

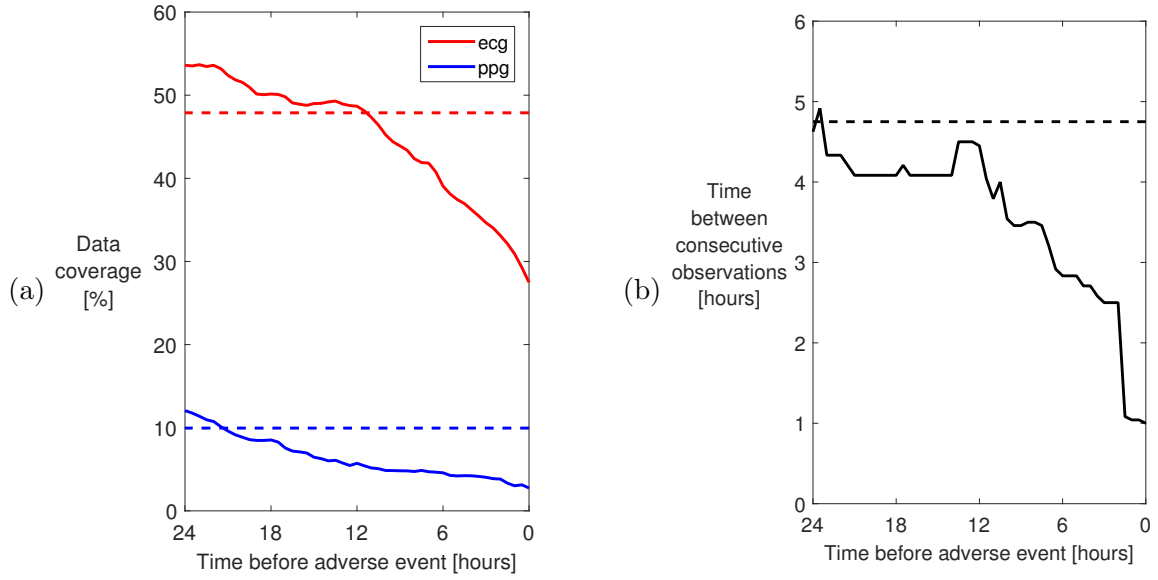

FIGURE 3.5: Changes in data coverage in the 24 hours before severe adverse events on the ambulatory ward: in (a) the coverage of continuous data decreased in the time before adverse events; whereas in (b) the coverage of intermittent physiological observation data increased in the time before deteriorations, as shown by a reduced time between consecutive observations. Dashed lines show median values for all other observations. Severe adverse events were defined as death, cardiac arrest and critical care readmission.

in the ambulatory setting. Of these, the data coverage rates for the ECG were substantially higher than those for the PPG. The numerics derived from the ECG and PPG signals (HR, PR and SpO_2) had similar data coverage rates to the signals from which they were derived. Secondly, in this setting the PPG was rarely acquired at times when the ECG wasn't acquired: the PPG was recorded for 31.4 (11.1 - 48.4) % of the time that ECG was recorded, and the ECG was recorded for 95.7 (87.7 - 99.0) % of the time that PPG was recorded. Therefore, any algorithms which can use either the ECG or the PPG, such as for estimation of HR, should prioritise using the ECG due to its improved data coverage. Furthermore, there would be little benefit to allowing an algorithm to switch between the ECG and PPG interchangeably, since this would only provide a minimal improvement in data coverage. Thirdly, the data coverage on the ambulatory ward varied with the progression of deteriorations: the data coverage of continuous data decreased as shown by a reduction in the proportion of time for which it was acquired (Figure 3.5(a)), whereas the frequency of intermittent physiological observations increased, as shown by a reduction in the time between observations (Figure 3.5(b)). This may reduce the advanced warning of deteriorations which can be provided by wearable sensors.

It is also helpful to consider how the window duration used for analysis of continuous signals affects the expected time between consecutive windows of continuous data. Most algorithms would benefit from a longer duration of data to provide more accurate parameter estimation.

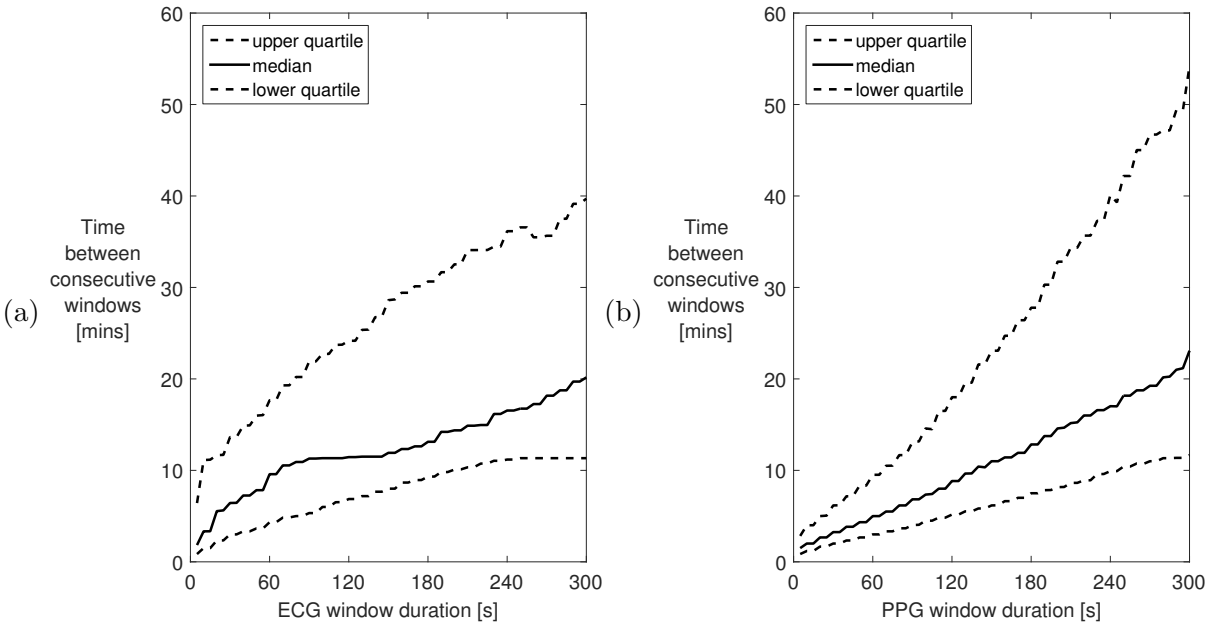

FIGURE 3.6: The impact of signal window duration on the frequency of available windows: (a) and (b) show how the time between consecutive windows of continuous ECG and PPG signals is impacted by the window duration. Gaps between segments of high quality continuous data mean that the time between consecutive high quality windows decreases as the duration of windows increases. This results in a compromise between the duration of data used in signal processing algorithms, and the frequency with which parameters can be estimated.

TABLE 3.5: The impact of signal window duration on the frequency of available windows. The median time between consecutive windows is provided for typical window durations used for signal processing algorithms. Results expressed as median (lower - upper quartiles).

Window duration [s]	Median time between consecutive windows [mins]	
	ECG	PPG
15	3.34 (1.50 - 11.17)	2.00 (1.18 - 4.00)
30	6.43 (3.00 - 13.63)	3.25 (2.00 - 6.17)
60	9.58 (4.33 - 17.67)	5.00 (3.00 - 9.52)
90	11.29 (5.33 - 21.91)	6.83 (4.02 - 13.08)
120	11.45 (6.87 - 24.12)	8.83 (5.17 - 18.00)
300	20.17 (11.33 - 39.71)	23.08 (11.67 - 54.23)

However, increasing the required window duration has the disadvantage of also increasing the time between consecutive windows, due to gaps in the data. Figures 3.6 (a) and (b) show how the time between consecutive windows increases with increased ECG and PPG window duration. It is clear that an increase in the window duration used by an algorithm will dramatically reduce the frequency with which parameters could be derived, since the time between consecutive windows is greater. Results for specific candidate window lengths are shown in Table 3.5. These

TABLE 3.6: Data coverage rates of intermittent physiological measurements. Definitions: *freq*: Frequency of measurements per day, median (lower - upper quartiles); *lag*: Time until next measurement [median hrs], median (lower - upper quartiles).

Variable	Entire Dataset		Critical Care		Ambulatory	
	freq	lag	freq	lag	freq	lag
HR	11.2 (9.4-13.4)	1.6 (0.6 - 4.4)	24.1 (22.7 - 25.1)	0.5 (0.3 - 0.8)	4.2 (3.7 - 4.9)	3.3 (1.6 - 5.7)
SpO_2	11.2 (9.2 - 13.3)	1.6 (0.6 - 4.3)	23.6 (22.1 - 24.7)	0.5 (0.3 - 0.8)	4.3 (3.8 - 5.0)	3.2 (1.5 - 5.6)
RR	11.2 (9.3 - 13.2)	1.6 (0.6 - 4.3)	23.9 (22.3 - 24.9)	0.5 (0.3 - 0.8)	4.2 (3.7 - 4.8)	3.2 (1.5 - 5.6)
SBP	8.8 (7.0 - 11.0)	2.6 (0.8 - 5.9)	18.4 (13.5 - 22.3)	0.7 (0.3 - 4.6)	4.3 (3.8 - 5.0)	3.2 (1.5 - 5.5)
DBP	8.8 (7.0 - 11.0)	2.6 (0.8 - 5.9)	18.4 (13.5 - 22.3)	0.7 (0.3 - 4.6)	4.3 (3.8 - 5.0)	3.2 (1.5 - 5.5)
temp	6.0 (5.2 - 7.0)	2.5 (1.1 - 4.6)	9.6 (8.3 - 11.3)	1.5 (0.7 - 2.7)	4.1 (3.6 - 4.6)	3.2 (1.5 - 5.6)
MAP	6.2 (4.3 - 8.3)	∞ (7.1 - ∞)	18.1 (13.3 - 22.1)	0.7 (0.3 - 35.9)	0.0 (0.0 - 0.1)	39.9 (14.4 - ∞)

results suggest that it may be possible to estimate parameters sufficiently frequently when using longer window durations. For instance, a window duration of 120 s would allow parameters to be estimated multiple times per hour, which is often enough to detect physiological changes indicative of deteriorations.

The data coverage rates of the intermittent physiological measurements are provided in Table 3.6. The five routinely measured vital signs (HR, SpO_2, RR, SBP and temp) were each measured approximately once every four hours on the ambulatory ward. DBP was also measured at the same frequency, although it may not be widely measured at other institutions since its inclusion in routine physiological measurements is not mandated [28]. MAP was not recorded on the paper charts, but only through the occasional electronic transmission of measurements to the electronic health record (EHR), explaining its low frequency of measurement. It may well be helpful to include the five routinely measured vital signs in algorithms for detection of deteriorations, since they were measured frequently enough to identify changes in physiology associated with deteriorations. Furthermore, the use of intermittent temp and SBP values may be particularly beneficial since these parameters are not available from the processed wearable sensor data.

The data coverage rates of the biochemistry results are provided in Table 3.7. The most fre-

TABLE 3.7: Data coverage rates of biochemistry results. Definitions:
freq: Frequency of measurements per day, median (lower - upper quartiles);
lag: Time until next measurement [median hours], median (lower - upper quartiles).

Variable	Entire Dataset		Critical Care		Ambulatory	
	freq	lag	freq	lag	freq	lag
Activated partial thromboplastin time	0.0 (0.0 - 0.0)	∞ (∞ - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)
Albumin	0.1 (0.0 - 0.2)	∞ (∞ - ∞)	0.5 (0.0 - 0.6)	∞ (∞ - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)
Bilirubin	0.0 (0.0 - 0.2)	∞ (∞ - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)
C-reactive protein	0.3 (0.1 - 0.4)	65.0 (28.4 - ∞)	0.3 (0.0 - 0.6)	54.0 (22.7 - 78.1)	0.2 (0.0 - 0.3)	∞ (34.5 - ∞)
Creatinine	0.8 (0.6 - 1.0)	20.1 (8.9 - 53.5)	1.5 (1.2 - 1.6)	11.5 (5.6 - 19.5)	0.4 (0.3 - 0.6)	35.8 (15.0 - ∞)
Glomerular filtration rate	0.8 (0.6 - 1.0)	20.1 (8.9 - 53.5)	1.5 (1.2 - 1.6)	11.5 (5.6 - 19.5)	0.4 (0.3 - 0.6)	35.8 (15.0 - ∞)
Haemoglobin	0.8 (0.6 - 1.0)	19.9 (9.1 - 54.7)	1.5 (1.2 - 1.7)	11.4 (5.4 - 19.5)	0.4 (0.3 - 0.6)	35.8 (15.0 - ∞)
Magnesium	0.5 (0.4 - 0.6)	∞ (17.2 - ∞)	1.4 (1.2 - 1.6)	11.8 (5.6 - 20.3)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)
Platelet count	0.8 (0.6 - 1.0)	19.9 (9.1 - 54.7)	1.5 (1.2 - 1.6)	11.4 (5.4 - 19.5)	0.4 (0.3 - 0.6)	35.8 (15.0 - ∞)
Potassium	0.8 (0.6 - 1.0)	20.5 (9.0 - 53.3)	1.4 (1.2 - 1.6)	11.6 (5.6 - 19.7)	0.4 (0.3 - 0.6)	35.6 (15.0 - ∞)
Sodium	0.8 (0.6 - 1.0)	20.1 (8.9 - 53.5)	1.5 (1.2 - 1.6)	11.5 (5.6 - 19.5)	0.4 (0.3 - 0.6)	35.8 (15.0 - ∞)
Red blood cell count	0.8 (0.6 - 1.0)	19.9 (9.1 - 54.7)	1.5 (1.2 - 1.7)	11.4 (5.4 - 19.5)	0.4 (0.3 - 0.6)	35.8 (15.0 - ∞)
Troponin-T	0.0 (0.0 - 0.0)	∞ (∞ - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)
Urea	0.2 (0.1 - 0.4)	∞ (∞ - ∞)	0.6 (0.3 - 1.0)	34.2 (11.6 - ∞)	0.0 (0.0 - 0.0)	∞ (∞ - ∞)
White blood cell count	0.8 (0.6 - 1.0)	19.9 (9.1 - 54.7)	1.5 (1.2 - 1.7)	11.4 (5.4 - 19.5)	0.4 (0.3 - 0.6)	35.8 (15.0 - ∞)

TABLE 3.8: Number of patients experiencing each of the most frequent types of adverse clinical event in the ambulatory setting, n (%)

Adverse event	Entire Dataset	Critical Care	Ambulatory
Atrial fibrillation	77 (34.1)	50 (22.1)	28 (12.4)
Cardiac pacing	107 (47.3)	84 (37.2)	26 (11.5)
Critical care readmission	11 (4.9)	0 (0.0)	11 (4.9)
Sepsis	26 (11.5)	17 (7.5)	9 (4.0)
Pleural drain	13 (5.8)	8 (3.5)	6 (2.7)
Cardiac failure	10 (4.4)	5 (2.2)	5 (2.2)
Return to surgery	18 (8.0)	15 (6.6)	4 (1.8)
Deep wound infection	6 (2.7)	2 (0.9)	4 (1.8)
Cardiac arrest	6 (2.7)	3 (1.3)	3 (1.3)
Altered mental state	21 (9.3)	18 (8.0)	3 (1.3)
Acute kidney injury	101 (44.7)	99 (43.8)	2 (0.9)

quently measured biochemistry values in the ambulatory ward were measured 0.4 (0.3 - 0.6) times per day, resulting in a median expected lag time until the next measurement of approximately 36 hours (1.5 days). Since deteriorations are expected to progress over periods of hours rather than days, routinely measured biochemistry values are unlikely to be beneficial for detection of deteriorations. Nonetheless, it has previously been observed that the use of biochemistry results can improve algorithms for detection of deteriorations [130].

Table 3.8 lists the the most common types of adverse clinical events in the ambulatory setting. The rate of adverse events in this patient population was low, as expected. The most common adverse events in the ambulatory setting were the onset of paroxysmal atrial fibrillation (occurring in 28 patients) and the need for cardiac pacing (26). However, neither of these are suitable candidates for predicting using multi-parameter wearable sensors since they can be readily detected from the ECG, and any precursory changes in physiology are likely to be confined to the ECG. In contrast, critical care readmission (11), sepsis (9), return to surgery (4), and cardiac arrest (3) are suitable candidates for prediction since they are more likely to be manifested across a range of physiological parameters. The low rates of adverse events mean that any predictive algorithms trained on this dataset should be hypothesis generating, and should be re-trained and validated on larger datasets. In addition, a technique such as cross validation should be used to maximise the contributions of the patients who experienced adverse events.

3.5 Dataset Evaluation

In this section the utility of the LISTEN Dataset for designing algorithms to detect deteriorations is evaluated.

The LISTEN Dataset is a valuable resource for assessing the utility of wearable sensors. It contains data collected from 226 patients, 200 of whom were provided with a wearable sensor for 3.7 (2.6 - 5.2) days. In contrast, similar datasets acquired in previous studies have been limited by either the wearable sensors being worn for short periods (typically 24 hours or less), or collecting data from a small number of patients (typically less than 50), as described in Section 2.4. Consequently, this dataset can be used for relatively robust analyses over a large number of patients, and over a length of time which is comparable to that of their expected usage.

The LISTEN Dataset contains all the variables which could be acquired routinely in the clinical setting in which the study was conducted, as detailed in Table 3.1. In addition to the routine monitoring, patients wore a wearable sensor for the entirety of their recovery on an ambulatory ward. In contrast, only a minority of patients wear a wearable sensor routinely. This facilitated capture of ECG and PPG signals throughout patients' hospital stays, making this a uniquely comprehensive record of the physiology of hospital patients recovering from cardiac surgery. In addition, all of the signals and parameters which are routinely monitored in critical care were captured. However, the dataset cannot be considered to be fully comprehensive since some other wearable sensors provide additional monitoring of activity, RR and temp, which are not available in this dataset. Nonetheless, it has the advantage that all of the variables acquired can be monitored in routine clinical practice, ensuring that any novel techniques developed based on continuous monitoring of these variables could be implemented in practice.

The inclusion of routinely acquired intermittent physiological measurements, and labels of adverse events, ensures that the dataset is highly suitable for developing and evaluating algorithms for detection of deteriorations. The intermittent measurements can be used to derive benchmark EWSs, allowing novel algorithms for detection of deteriorations to be compared to current practice. In addition, the labels of adverse events can be used to classify data from each patient to train algorithms.

The data coverage rates are those which can be expected in a real-world clinical setting (Table 3.4). The highest data coverage rate was given by the ECG in both critical care and ambulatory

settings. Therefore, it should be the primary choice of signal for analyses which can be performed using several signals, such as RR estimation and heart rate variability (HRV) analysis, which could use either the ECG or the PPG. The PPG signal had a relatively low data coverage rate, which should be considered when evaluating the usefulness of deriving novel parameters from the signal. The data coverage rates tended to be much higher in critical care. Therefore, it may be advantageous to develop techniques on the critical care data, and use the ambulatory data to test their performance.

The LISTEN Dataset has been pre-processed using several trial-specific processing procedures, ensuring that future users can begin analyses quickly without having to perform pre-processing to clean up the data. Four continuous signals were quality assessed, reducing the inaccuracies in the numerics derived from them. This is particularly important in the ambulatory setting where motion artifact and poor sensor contact are common. A novel SQI was developed for use with the ImP signal, allowing users to have confidence in both the RRs and the timings of individual breaths derived from this signal. This facilitates analyses which require accurate breath timings, such as respiratory rate variability analyses [131], and those which require precise RRs, such as the use of ImP-derived RRs as a reference for the assessment of algorithms to estimate RR from the ECG and PPG [4]. In contrast, other comparable datasets containing continuous physiological data, such as the MIMIC Database acquired using static monitors in critical care [132], require substantial pre-processing before analyses can be conducted. This makes the LISTEN Dataset ideal for rapid analyses without the need for considerable pre-processing.

The data acquisition procedures used to collect the LISTEN Dataset were not perfect, leading to missing, and in some cases, erroneous data. Approximately one third of the patients who wore a wearable sensor asked to have it removed before the end of their hospital stay [123]. In addition, technical malfunctions of the data acquisition system led to gaps in the continuous monitoring data of several hours or even days on occasion. Furthermore, the intermittent physiological measurements were acquired by transcription from handwritten charts, which were sometimes illegible. Data gaps such as these mean that some subjects may not contribute data to particular analyses. Future users should also be aware of the limitations of the monitoring equipment and data acquisition software. These included: (i) sampling frequencies of 125 Hz, potentially inhibiting particular analyses such as HRV analyses; (ii) varying, unknown timing offsets between the ECG and PPG signals, inhibiting pulse transit time calculations; and, (iii) the acquisition of signals using the amplitude scaling applied by the ward staff, potentially

reducing the resolution of signals. Nonetheless, the acquired signals and parameters represent those which can be reasonably collected during current clinical practice.

The patient population for this trial was chosen to facilitate relatively high data coverage rates. However, there were several disadvantages to using this population. Firstly, the group of subjects was relatively homogeneous since each patient's critical illness was caused by cardiac surgery. In contrast, medical inpatients can be admitted because of a range of illnesses, meaning that any findings may not be generalisable beyond cardiac patients. However, findings may still have ramifications since over 30,000 patients undergo cardiac surgery in England and Wales each year [120]. Secondly, since of all the subjects had undergone cardiac surgery, their physiology may have been deranged due to the after-effects of the operation, making it difficult to differentiate between derangements indicative of deteriorations, and those due to surgery. Thirdly, all physiological measurements acquired were available to clinicians in real-time, so would have prompted changes in treatment thus altering both physiology and end-points from their natural courses. The impact of this could have been reduced by selecting a patient population which does not receive continuous ambulatory monitoring as part of routine care.

3.6 Final Remarks

This chapter has described the processes for, and results of, assembling a physiological dataset with which to test the hypothesis of this thesis. The LISTEN Dataset is a highly comprehensive physiological dataset containing data acquired from patients recovering from major cardiac surgery in hospital. The data is presented as two datasets: the *Prepared Dataset*, which is designed to contain all of the data acquired, after elimination of nonsensical data; and the *Processed Dataset*, which has been subject to quality assessment. The *Processed Dataset* is recommended for most analyses, unless real-world data is required, in which case the *Prepared Dataset* should be used.

The LISTEN Dataset contains continuous physiological monitoring data acquired from critical care using static bedside monitors, and the ambulatory setting using wearable sensors. Continuous ECG and PPG signals are available throughout, with additional ImP and ABP signals in critical care. In addition, intermittent physiological measurements, biochemistry results, demographics and labels of adverse events are provided. Furthermore, EWSs can be derived from the intermittent measurements, acting as benchmark algorithms for detection of deteriorations.

Initial analyses of the dataset provided guidelines for the design of algorithms to detect deteriorations when using this dataset:

- Only ECG and PPG signals were available from wearable sensors. Therefore, algorithms for use in the ambulatory setting should use these signals and the numerics derived from them (HR, PR and SpO_2).
- The ECG was captured for a greater proportion of the time than the PPG. Furthermore, the PPG was only rarely captured at times when the ECG was not, meaning that the ECG should preferably be used for algorithms where possible.
- Data coverage rates of continuous monitoring data decreased before severe adverse events, which may impair the detection of deteriorations.
- The use of shorter windows (*e.g.* 30 s) of signals to estimate additional parameters, such as RR, is beneficial for increasing the frequency of available windows. However, the use of longer windows (*e.g.* 120 s) would still allow additional parameters to be estimated multiple times per hour, potentially providing early detection of deteriorations.
- The five routinely measured vital signs (HR, RR, SpO_2, BP and temp) were the only intermittent measurements which were measured frequently enough to be expected to be of use in algorithms for detection of deteriorations.
- The adverse events which occurred most often in the dataset, and which are expected to be manifested across multiple physiological parameters are: cardiac arrest, sepsis, readmission to critical care, and return to surgery.

The LISTEN Dataset has the advantage of being a real-world dataset, although its generalisability to other acutely-ill patient populations is potentially limited. It has the advantage that it contains variables which can be routinely acquired from hospital patients. This means that any algorithms developed using the dataset could be applied in current practice. Furthermore, since the data was collected across long monitoring periods from a high number of patients, the data coverage rates are those which could be expected in clinical practice. However, the dataset has the disadvantage that it contains data solely from patients recovering from cardiac surgery. Therefore, any findings may not be generalisable beyond this population. In addition, this population suffered from a relatively low rate of adverse events, meaning algorithms developed on this dataset may need to be re-trained and validated on other datasets.

Chapter 4

Estimating Respiratory Rate from the Electrocardiogram and Photoplethysmogram: a Systematic Review and Toolbox of Resources

This chapter presents a systematic review of algorithms to estimate respiratory rate (RR) from the electrocardiogram (ECG) and photoplethysmogram (PPG) signals. It starts by describing the rationale for estimating RR from the ECG and PPG. This is followed by an explanation of the physiological basis for estimating RR from the ECG and PPG. The methodology for a systematic review of the literature is then presented. The results of this review are presented focusing on the systematic search process, the characteristics of the included publications, the techniques used for RR algorithms, and the methodologies used to assess RR algorithms. A discussion provides a critical review of the state of the literature, culminating in the identification of key limitations which should be addressed in future research. Finally, a toolbox of resources to aid research into RR algorithms is presented consisting of: (i) novel datasets for assessment of RR algorithms, (ii) a novel signal quality index (SQI) for use with the impedance pneumography (ImP) signal, and (iii) a set of benchmark RR algorithms. These resources are being made publicly available at: http://peterhcharlton.github.io/RRest.

4.1 Rationale

RR is an important physiological parameter which provides valuable diagnostic and prognostic information. It is frequently measured in acutely-ill hospital patients since it is a highly sensitive marker of deteriorations, as detailed in Sections 2.2 and 2.5. It is also widely used in many other clinical scenarios, such as: in emergency department screening of both children [133] and adults [134]; in the identification of pneumonia [135] and sepsis [136]; as a marker of hypercarbia [104] and pulmonary embolism [137]; and, for prognosis in the ICU [138]. RR is usually measured by manually counting chest wall movements, which is time-consuming, inaccurate [32], and poorly carried out [101, 102, 104], as described in Section 2.5. In addition, RR monitoring is not widely incorporated into wearable sensors. Therefore, there is an unmet clinical need for an automated, electronic method for measuring RR. If an unobtrusive method could be developed then it may have utility both for taking intermittent measurements, and for continuous monitoring using wearable sensors.

A potential solution is to estimate RR from a non-invasive signal which is modulated by respiration and is easily measured. Two such signals are considered in this review: the ECG and the PPG. The ECG is measured for a short period of time during screening for heart disorders. It is also monitored continuously in high acuity settings such as the intensive care unit or coronary care unit, and in hospital patients at risk of dysrhythmias [56]. In addition, wearable sensors often monitor the ECG. Indeed, some recently developed sensors incorporate ECG monitoring into a chest-worn patch, providing continuous single-lead ECG monitoring over a period of several days in a highly unobtrusive manner [81]. The PPG is routinely measured in a wide range of clinical settings to obtain arterial blood oxygen saturation (SpO_2) and pulse rate (PR) measurements. In addition, the PPG is continuously monitored in critically-ill patients, and can also be monitored continuously in ambulatory patients using wearable sensors. Furthermore, non-contact video-based technology is being developed for continuous PPG monitoring in a wide range of clinical scenarios, without the need for any equipment to be attached to the patient [139]. Therefore, both the ECG and PPG signals are widely available. In addition, they are both modulated by respiration, making them a highly attractive means of obtaining unobtrusive, electronic RR measurements. Furthermore, they were both acquired throughout the LISTEN Trial (Section 3.4), meaning that techniques for estimation of RR from the ECG and PPG could be used to derive RR from the continuous data in the LISTEN dataset.

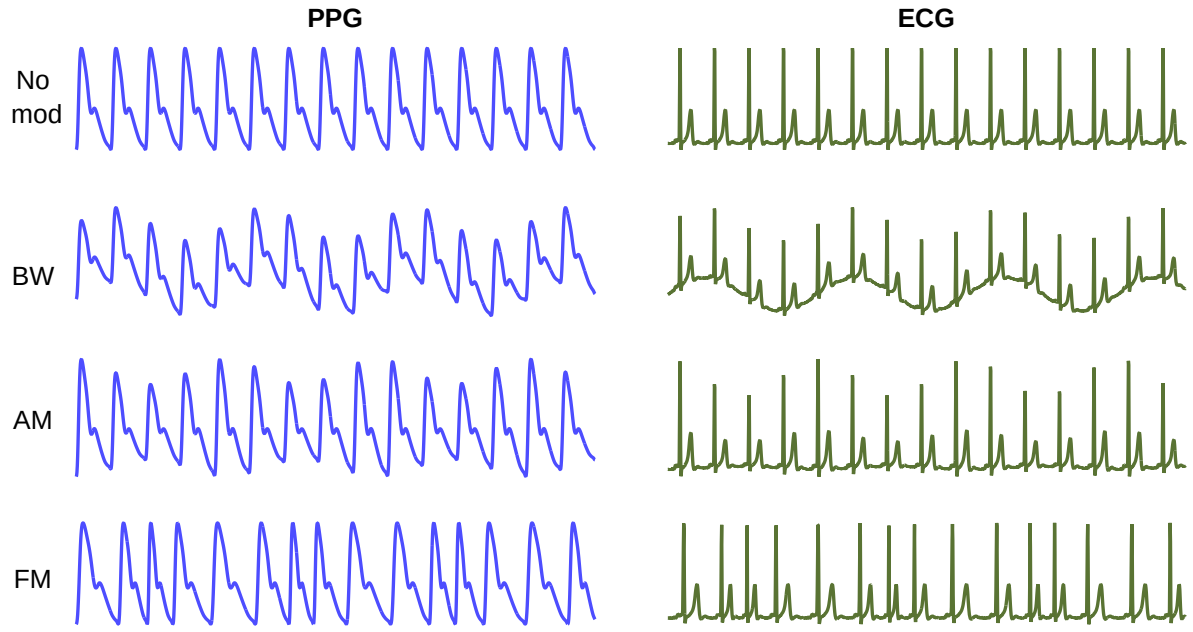

FIGURE 4.1: Respiratory modulations of the ECG and PPG: Both the ECG and the PPG are subject to three respiratory modulations: baseline wander (BW), amplitude modulation (AM) and frequency modulation (FM). Source: [10] (CC BY 3.0, DOI: 10.1088/0967-3334/37/4/610)

Over 100 algorithms have been proposed to estimate RR from the ECG or PPG [10]. However, the literature surrounding these algorithms has not been synthesised. Consequently, it is difficult to identify all of the algorithms which have been proposed, and to determine which, if any, RR algorithm is most suitable for clinical use.

A systematic review was undertaken to establish: (i) the techniques used to estimate RR; (ii) the methodology used to assess RR algorithms; and, (iii) the reported performances of RR algorithms.

4.2 Physiological Basis for Respiratory Rate Algorithms

Algorithms to estimate RR from the ECG and PPG are based on extraction of one or more respiratory modulations from the signals. Both the ECG and PPG exhibit three types of respiratory modulation as illustrated in Figure 4.1: baseline wander (BW), amplitude modulation (AM) and frequency modulation (FM). RR can be estimated from the ECG or PPG by extracting a signal which is dominated by one or more of these modulations. For instance, a low-pass filter could be used to extract the low frequency respiratory modulation due to BW. The resulting signal could then be used to estimate RR through spectral analysis.

TABLE 4.1: Physiological mechanisms of ECG and PPG respiratory modulation. For a comprehensive treatment see [79] and [140]. Source: [7] (CC BY 3.0, DOI: 10.1088/1361-6579/aa670e)

Modulation	ECG	PPG
BW	Beat morphology is influenced during respiration by two mechanisms: (i) changes in thoracic impedance, and (ii) changes in the orientation of the electrical axis of the heart relative to ECG electrodes [79].	Changes in tissue blood volume caused by: (i) transmitted changes in intrathoracic pressure; and (ii) vasoconstriction of arteries during inhalation, transferring blood to more central veins [144].
AM		Stroke volume is reduced during inhalation due to changes in intrathoracic pressure, affecting pulse amplitude [140].
FM	FM is the manifestation of respiratory sinus arrhythmia (RSA) [145] which causes heart rate to increase during inspiration and decrease during exhalation. It is caused by at least three mechanisms: (i) changes in intrathoracic pressure during inhalation stretch the sino-atrial node, increasing heart rate (HR); (ii) increased vagal outflow during exhalation reduces HR; and (iii) reduced intrathoracic pressure during inhalation decreases left ventricular stroke volume, causing a baroreflex-mediated increase in HR [146].	

Several physiological mechanisms are responsible for the respiratory modulations of the ECG and PPG. They have been reported previously in [79] and [140], and are summarised in Table 4.1. The mechanisms which cause BW and AM differ between the ECG and PPG, whereas a single mechanism causes FM in both signals. The mechanisms are not fully understood. However, it is clear that the strength of any particular respiratory modulation can be affected by a range of factors. For instance, respiratory sinus arrhythmia (RSA, which causes FM) and chest wall expansion (which is linked to BW and AM) both diminish with age [141], [142], [143], [10]. The strength of a modulation may affect the reliability of RR estimates which can be derived from that modulation.

4.3 Methods

4.3.1 Approach

Publications containing algorithms for estimation of RR from the ECG and PPG were identified systematically using the following methodology. Firstly, it was decided that publications would

be eligible for inclusion if they described: (i) an algorithm which operated on either the single-lead ECG signal or the single-channel PPG; and (ii) an algorithm for estimation of either RR or breath-to-breath intervals. Publications were excluded if they met one of the following exclusion criteria: (i) they were written in a language other than English; (ii) no full text was available (after contacting authors where possible); or, (iii) they were patents. Secondly, an initial, pragmatic search was performed for publications containing algorithms meeting these criteria. Thirdly, the titles of the publications identified during the pragmatic search were analysed to determine a search strategy for use in a systematic search. Fourthly, a systematic search was conducted using the following online databases: PubMed, Scopus, IEEE Xplore, and Google Scholar. Finally, publications containing algorithms meeting the inclusion criteria were identified through screening of titles, followed by full publications.

4.3.2 Determining a search strategy

The initial pragmatic search yielded a total of 90 qualifying publications. This allowed a search strategy to be determined as follows. Inspection of the publication titles revealed three common themes: (i) the process of respiration; (ii) a mathematical process; and, (iii) description of the input signal. Words which occurred in at least 5% of the publication titles were sorted into one of these three categories, or deemed to be irrelevant. This yielded 27 keywords: three in the respiration category; 15 in the mathematical process category; and nine in the input signal category. A total of 76.7% of the titles contained at least one of the keywords from each category. Therefore, a search strategy was devised consisting of three search terms (one for each category), in which publication titles had to contain at least one keyword from each search term to be included. The next step was to eliminate keywords which did not add value to the search strategy, to increase the specificity of the search. A total of 10 keywords did not add value to the search strategy, as demonstrated by there being no reduction in the sensitivity of the search strategy (from 76.7%) when each was individually omitted. Omission of all 10 only slightly reduced the sensitivity of the search strategy to 75.6%. Therefore, this was chosen as the search strategy to be used in a systematic search. The final keywords are listed in Table 4.2. These were used to perform a systematic search on the 12^{th} and 13^{th} August 2016. Both publications identified through the systematic search and those identified through the initial pragmatic search were screened against the inclusion and exclusion criteria.

TABLE 4.2: Systematic review search terms. To be included, publication titles had to contain at least one keyword corresponding to each search term.

Search term	Theme	Keywords
S1	Respiration	breathing, respiration, respiratory
S2	Mathematical process	derivation, derived, estimation, extraction, methods, rate, rates
S3	Input signal	ECG, electrocardiogram, photoplethysmogram, photoplethysmographic, photoplethysmography, PPG, pulse

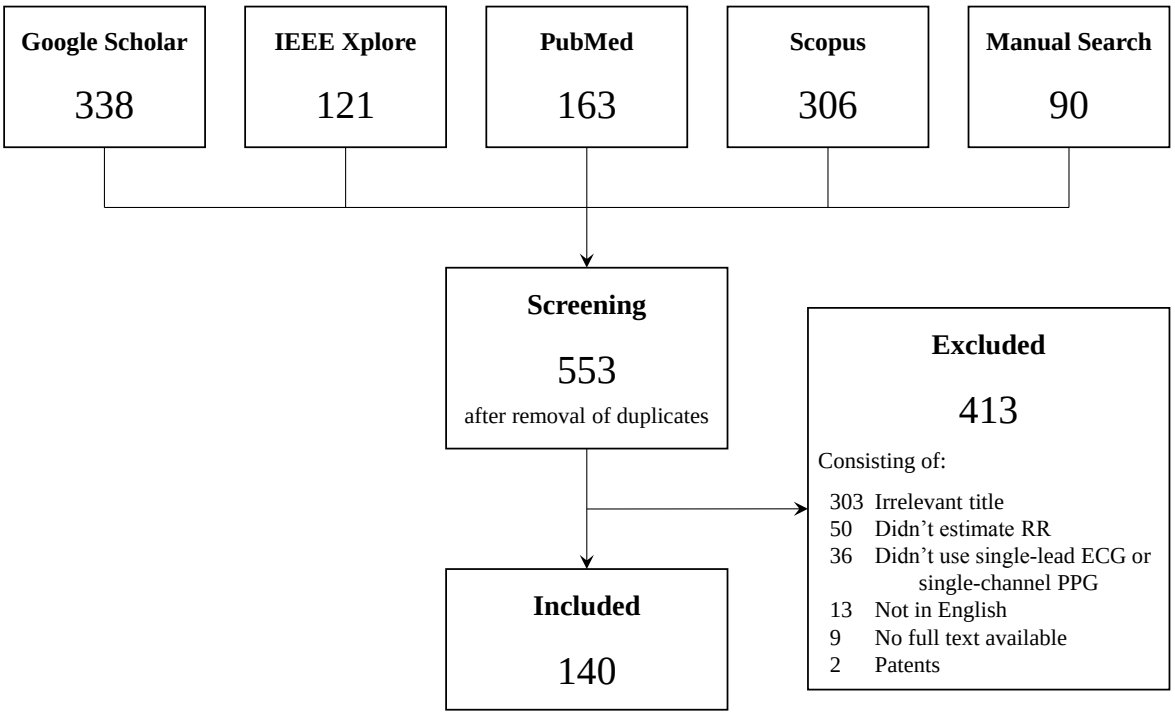

FIGURE 4.2: Identification and screening of publications describing respiratory rate (RR) algorithms showing: the number of publications identified by the searches, screened, excluded and included in the final analysis.

4.4 Results

4.4.1 Search and screening

The results of the identification and subsequent screening of publications are summarised in Figure 4.2. The search identified 553 publications. These were screened against the inclusion and exclusion criteria, leading to exclusion of 413 publications. Most excluded publications were

TABLE 4.3: The utility of databases searched in a systematic review of respiratory rate (RR) literature. Definitions: No. pubs - number of publications included in the final analysis; No. unique pubs - number of publications included in the final analysis which were unique to this database; Sen - sensitivity; PPV - positive predictive value.

Database	No. pubs	No. unique pubs	Sen [%]	PPV [%]
Google Scholar	36	7	25.7	10.7
IEEE Xplore	59	2	42.1	48.8
PubMed	48	2	34.3	29.4
Scopus	91	4	65.0	29.7
Manual Search	90	31	64.3	100.0

deemed to be irrelevant based on their titles. In addition, many articles were excluded because they did not present algorithms to estimate RR. These articles often presented methods for extracting a respiratory signal from the ECG or PPG, without estimating RR from that signal. A total of 140 publications were included in the final analysis: [10, 80, 108–113, 147–278]

Statistics describing the utility of each database are presented in Table 4.3. The most comprehensive database was Scopus, which contained nearly two thirds of the publications included in the final analysis (shown by its sensitivity of 65.0 %). Approximately one third of the results retrieved from Scopus were included in the final analysis (shown by its positive predictive value, PPV, of 10.7 %). The least comprehensive database was Google Scholar, which returned only 25.7 % of the included publications, and only approximately one tenth of its results were included (shown by its PPV of 10.7 %). It is notable that the manual search strategy missed 35.7 % of the publications included in the final analysis, and the automated search strategy consisting of the four electronic databases missed 22.1 % of the included publications.

4.4.2 Characteristics of included publications

The included publications were published between 1991 and 2016. Only three were published in the seven years following the initial publication in 1991. Since 1999, the rate of publication has risen steadily up to the present rate of approximately 16 publications per year, as shown in Figure 4.3. Approximately half of the included publications (72 out of 140, 51.4 %) were presented at conferences. Most of the remaining publications were journal articles (63, 45.0 %). The remainder were theses (5, 3.6 %).

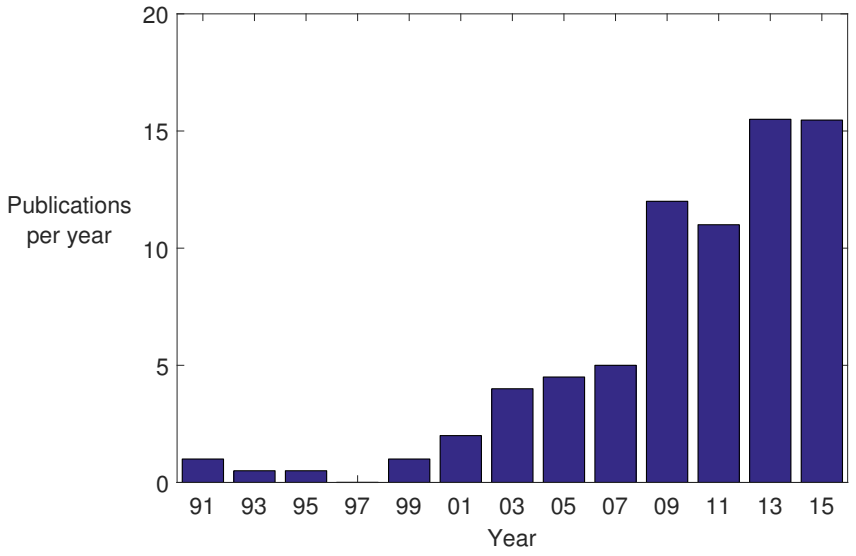

FIGURE 4.3: The number of publications describing respiratory rate (RR) algorithms published each year: This histogram illustrates how the number of publications describing RR algorithms each year since the first publication in 1991. The rate of publication has risen steadily since 1999.

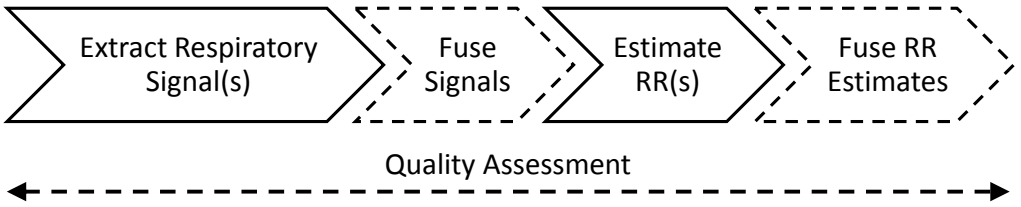

FIGURE 4.4: The stages of a respiratory rate (RR) algorithm. Dashed stages are optional.

4.4.3 Respiratory rate algorithms

During the review of RR algorithms it was observed that they consisted of up to five stages, as illustrated in Figure 4.4. The role of each stage is as follows. *Extract Respiratory Signal(s)* consists of extracting one or more time series dominated by respiratory modulation. *Fuse Signals* is an optional stage in which multiple respiratory signals can be fused to give one respiratory signal in either the time or frequency domain. *Estimate RR(s)* consists of estimating a RR from a window of a respiratory signal. *Fuse RR Estimates* is an optional stage used to fuse multiple RR estimates to obtain one final estimate. *Quality Assessment* can be used at various stages throughout an algorithm to reject or mitigate against imprecise estimates. The signal processing steps which have been used in the literature at each stage are now described.

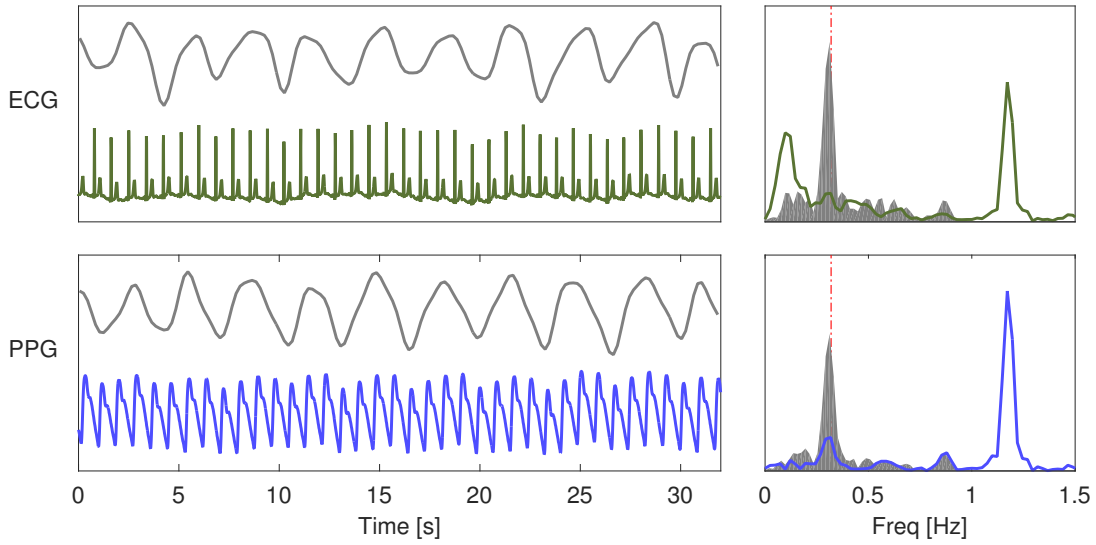

FIGURE 4.5: Extraction of a respiratory signal from the ECG or PPG: ECG (upper plot) and PPG (lower plot) signals and extracted respiratory signals (grey) are shown on the left. The corresponding frequency spectra are shown on the right. The frequency spectra of the raw ECG and PPG signals are dominated by the cardiac frequency content at 1.2 Hz. In contrast, the frequency spectra of the extracted respiratory signals are dominated by respiration at 0.3 Hz, which is approximately the RR provided by simultaneous ImP measurement (shown by the dashed line). Respiratory signals were extracted from amplitude modulation (AM) of the ECG and PPG.

4.4.3.1 Extracting respiratory signals

ECG and PPG signals are primarily cardiac in origin, with secondary respiratory modulations of much lower magnitudes. Therefore, the first stage of an RR algorithm is extraction of a respiratory signal, *i.e.* a time series dominated by respiratory modulation, from which RRs can be much more easily estimated. Figure 4.5 shows examples of respiratory signals extracted from the ECG and PPG. The frequency spectra demonstrate that the extracted respiratory signals are dominated by respiratory modulation. Techniques for extraction of a respiratory signal fall into two categories: either feature-based or filter-based techniques. Feature-based extraction of a respiratory signal consists of the extraction of a time series of beat-by-beat feature measurements (such as the amplitude of each QRS complex). Filter-based extraction consists of filtering the raw signal to attenuate non-respiratory frequency components (such as band-pass filtering the PPG to extract a respiratory signal exhibiting BW). The individual processing steps used for each approach are shown in Figure 4.6, and are now described.

Processing steps common to both feature- and filter-based techniques

The first and final steps of the extraction of respiratory signals are common to both feature- and filter-based techniques. The first step is the elimination of very low frequency (VLF) components

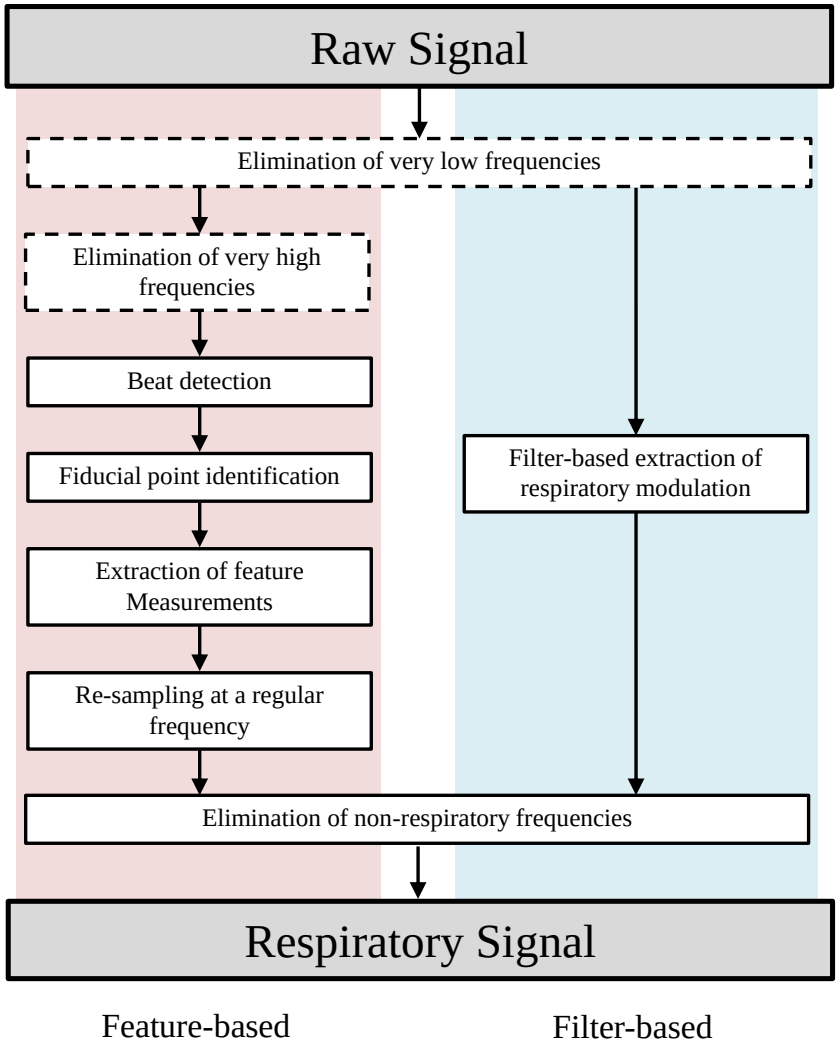

FIGURE 4.6: Processing steps for extraction of a respiratory signal: The steps required to extract a respiratory signal using the two possible approaches: (a) Feature-based respiratory signals are extracted by constructing a time series of beat-by-beat feature measurements (on left); (b) Filter-based respiratory signals are extracted by filtering the raw signal to produce a signal in which the non-respiratory frequency components of the raw signal are attenuated (on right). Dashed lines indicate optional steps.

of the PPG and ECG, *i.e.* those at sub-respiratory frequencies. VLFs are eliminated using a finite-impulse response high-pass filter such as: a median filter [168, 240, 244]; subtraction of a baseline trend calculated using a linear or polynomial fit [177, 226], or using measurements of the baseline at a specific point in the cardiac cycle (such as the baseline shortly before the QRS complex [225], or the midpoints between successive R waves [150] in the ECG). A cut-off frequency of 0.03-0.05 Hz is typically chosen [226, 256, 257, 264]. This stage is not required if the ECG or PPG monitoring device removes VLFs prior to output. The subsequent processing steps for extraction of respiratory signals then differ depending on whether a feature- or filter-based technique is being used, until the final step which is common to all techniques. The final step is

the elimination of frequency components outside of the range of plausible respiratory frequencies. Band-pass filtering is used to eliminate non-respiratory frequencies with cut-offs at either end of the range of plausible respiratory frequencies [113, 153, 168, 198, 238]. Several ranges of plausible respiratory frequencies have been proposed in the literature, with no consensus on the optimal range. Furthermore, the range may need to be adjusted according to the patient population, particularly for paediatrics [279]. Indeed, some algorithms use a range with varies according to age [280], or a range which adapts to the HR [108, 231, 232, 267, 281]. As a guideline, Karlen *et al.* used a conservative range of 4 - 65 breaths per minute (bpm) [80]. The steps of elimination of VLFs before extraction of a respiratory signal, and elimination of non-respiratory frequencies after extraction of a respiratory signal, help to avoid erroneously identifying spurious frequency content as the RR. The intermediate steps for extraction of a respiratory signal differ according to whether a feature- or filter- based technique is being used, and are now described.

Filter-based techniques for extraction of respiratory signals

Filter-based techniques for extraction of a respiratory signal are performed in a single step. Several techniques have been proposed, as listed in Table 4.4.

Processing steps specific to feature-based extraction of respiratory signals

Feature-based extraction of respiratory signals consists of five processing steps. Three of these steps are common to all feature-based techniques, with the two remaining steps changing according to the technique being used.

The first of these five steps is the elimination of very high frequency (VHF) noise by low-pass filtering to improve the accuracy of beat detection and feature measurements. Higher cut-off frequencies are used for the ECG (*e.g.* 40, 75, or 100 Hz [159, 187, 257]) than the PPG (*e.g.* 10 or 35 Hz [113, 187]), to retain the high frequency content of the QRS complex. In addition, the ECG is particularly susceptible to power-line interference [283], which may be eliminated using an additional band-stop filter. Whilst a fixed filter is commonly used, an adaptive filter (which adapts to the noise content) has also been used [198, 284, 285]. Commercial monitoring devices typically remove VHFs prior to signal output, similarly to VLFs.

Next, feature-based techniques require the detection of individual beats. ECG beat detection is typically performed by identifying QRS complexes, from which the R-waves can be identified. Approaches for identifying QRS complexes include: threshold beat detection [187]; adaptive

TABLE 4.4: Filter-based techniques for extraction of respiratory signals. Techniques used in this thesis are indicated by *

Abbr.	Technique
X_{A1}*	BW: Band-pass filter to eliminate frequencies outside the range of plausible respiratory frequencies using a digital filter [214].
X_{A2}*	AM: The maximum amplitude of the continuous wavelet transform within plausible cardiac frequencies (approx. 30-220 beats per minute) [282].
X_{A3}*	FM: The frequency corresponding to the maximum amplitude of the continuous wavelet transform within plausible cardiac frequencies [282].
X_{A4}*	BW, AM, FM: Filter using the centred-correntropy function [110].
X_{A5}	Decimation, consisting of detrending the signal, low-pass filtering to eliminate frequencies higher than respiration, and re-sampling at a reduced sampling frequency [108] of 1 - 2 Hz [177].
X_{A6}	AM, FM: Estimate instantaneous amplitudes and frequencies using the Teager-Kaiser energy operator [263].
X_{A7}	Use empirical mode decomposition to identify the oscillation modes within a signal, and extract a respiratory signal as the sum of the IMFs indicative of respiration [254].
X_{A8}	FM: Calculate instantaneous frequencies using variable frequency complex demodulation [164].
X_{A9}	AM, FM: Decompose signal using the discrete wavelet transform to reconstruct the detail signal at a predefined decomposition scale [157].
X_{A10}	Extract respiratory oscillation using principal component analysis by detecting periodicity using singular value decomposition to identify periodicity [222].

threshold beat detection using the first derivative of the ECG [72, 73]; a combination of methods using adaptive threshold detection and the curve length transform [111]; and, wavelet transform methods [286]. The adaptive threshold, curve length and wavelet methods have all been evaluated on the benchmark MIT/BIH arrhythmia database, achieving sensitivities and PPVs for beat detection of over 99.5%, as described in [72, 73], [287] and [286, 288, 289] respectively. Once QRSs have been detected, R-waves can be identified. They have been identified as the maximum ECG value between consecutive QRS onsets [191], or more specifically, as the maximum value within a certain time of the detected QRS (*e.g.* within 20 ms [159]). PPG beat detection is similarly performed by identifying cardiac pulses, from which pulse peaks can be identified. Methods of identifying cardiac pulses include: band-pass filtering to eliminate non-cardiac frequencies (*e.g.* 0.5-2.0 Hz), followed by peak detection [186]; threshold beat detection using the first derivative of the PPG [250]; the Incremental-Merge Segmentation (IMS) algorithm [74, 80]; and, identifying regions for pulse peaks using a simultaneous ECG signal

TABLE 4.5: Feature-based techniques for extraction of respiratory signals. Techniques used in this thesis are indicated by *

Abbr.	Technique
X_{B1}*	BW: mean amplitude of troughs and proceeding peaks [10].
X_{B2}*	AM: difference between the amplitudes of troughs and proceeding peaks [80].
X_{B3}*	FM: time interval between consecutive peaks [237], [80].
X_{B4}*	BW: mean signal value between consecutive troughs [112].
X_{B5}*	BW, AM: peak amplitude [80].
X_{B6}*	BW, AM: trough amplitude [112].
X_{B7}*	FM: QRS duration [244]. Q and S waves were identified as the minima immediately before and after the R wave [112].
X_{B8}*	AM, FM: QRS area [256], defined as the integral of the ECG between Q and S waves after subtraction of a baseline linearly interpolated between Q and S waves.
X_{B9}*	BW: Kernel principal component analysis using a radial basis function, with the variance of the Gaussian kernel determined by maximising the difference between the first eigenvalue and sum of the remainder [291].
X_{B10}*	FM: PPG pulse width estimated using a wave boundary detection algorithm [199].
X_{B14}	Calculate the kurtosis between adjacent peaks [174].
X_{B15}	Difference between durations of the upslope and downslope of the PPG [265].

[113, 198, 236, 241, 290]. Pulse peaks are then identified as the maximum PPG value between consecutive cardiac pulse onsets [186, 191].

The third and fourth steps are specific to the particular feature-based technique being used. Many techniques require identification of additional fiducial points beyond simply R-waves and pulse peaks. These points are usually identified using heuristic rules. Once the required fiducial points have been identified, measurements of features which vary with respiration are obtained from the time-amplitude co-ordinates of the fiducial points. Several features have been used, as summarised in Table 4.5.

The fifth step specific to feature-based extraction of respiratory signals is to re-sample the respiratory signal at a regular sampling frequency. This is necessary since signals obtained from beat-by-beat feature measurements are irregularly sampled (once per beat), whereas the subsequent elimination of non-respiratory frequencies is conducted using filtering techniques which require a regularly sampled signal. Often linear [80, 268] or cubic spline interpolation [198] is used. More complex methods include: Berger's algorithm, designed for use with an

FM signal [292], and used in [80, 191]; the Integral Pulse Frequency Modulation model, also designed for use with FM signals [293], and used in [254]; and, the discrete wavelet transform [244]. Respiratory signals are typically re-sampled at 4-10 Hz, completing the processing steps specific to feature-based extraction of respiratory signals.

Additional approaches

Two additional approaches have been proposed which are not based on the physiological mechanisms described earlier in Section 4.2: the use of pulse transit time (PTT) and the electromyogram (EMG). PTT can be estimated from the time interval between the R-wave and subsequent PPG pulse onset [149, 187], or between the R-wave and the time at which the PPG pulse has risen by 50% of its amplitude [198]. It has been suggested that this varies with respiration, possibly due to changes in pulse wave velocity induced by respiration [294]. It has been suggested that an EMG signal can be extracted from the high frequency content ($>$ 250 Hz) of the ECG exhibiting the activation of the diaphragm and intercostal muscles during respiration [185]. However, neither of these signals can be extracted from the LISTEN Dataset described in Chapter 3, due to the low sampling frequency (125 Hz) of its signals, and the changing time alignment between the ECG and PPG. Therefore, they are not considered any further.

4.4.3.2 Fusing respiratory signals

The second stage of RR algorithms is the fusion of multiple respiratory signals to provide one respiratory signal from which RR can be estimated. Multiple respiratory signals arise in two scenarios. Firstly, multiple respiratory signals can be extracted simultaneously from one patient, either by using both the ECG and PPG, or by using multiple methods for extracting respiratory signals. Secondly, a respiratory signal can be segmented into several (often overlapping) time periods, and the signals within each time period treated as individual respiratory signals. In each scenario, fusion is conducted in either the time or the frequency domain, depending on the domain in which the subsequent RR estimation component functions. Methods for performing fusion of multiple respiratory signals are listed in Table 4.6. This stage is optional, and is one of two ways in which multiple respiratory signals can be used to increase the robustness of RR estimates, the other being the fusion of RR estimates derived from multiple signals.

TABLE 4.6: Techniques for fusion of respiratory signals

Abbr.	Technique
	Frequency domain, M_{A1} to M_{A6}
M_{A1}	Spectral-averaging [256]: Calculate the individual power spectra of multiple respiratory signals, and then find the average power spectrum.
M_{A2}	Peak-conditioned spectral-averaging [113, 198]: only spectra which are sufficiently *peaked* are included in calculation of the *peak-conditioned* average power spectrum. To qualify: (i) a spectrum must contain a peak within a confidence interval around the previous RR estimate with a magnitude $\geq$ 85% [113] (formerly $\geq$ 75% [198]) of the largest peak in the spectrum; (ii) at least a certain proportion of the power within this confidence interval must lie in a narrower interval centred on the largest peak within this confidence interval.
M_{A3}	Cross time-frequency analysis [236]: Use the smoothed pseudo Wigner-Ville distribution to estimate time-frequency spectra between two signals.
M_{A4}	Time-frequency coherence [236]: used to measure the degree of coupling between two signals.
M_{A5}	Vector autoregressive (AR) modelling [217]: The poles of multiple AR models (one for each respiratory signal) are calculated. Only those poles which are common to both models, and which fall within the range of plausible respiratory frequencies, are used to extract a respiratory signal.
M_{A6}	Temporal spectral averaging [113, 198]: Calculate an averaged frequency spectrum from overlapping periods of a respiratory signal.
	Time domain, M_{B1} and M_{B2}
M_{B1}	Point-by-point multiplication [197].
M_{B2}	Use of a neural network to identify periods of inhalation and exhalation [149, 186].

4.4.3.3 Estimating respiratory rate

The third component of RR algorithms is the estimation of RR. The input to this component is a single respiratory signal, and the output is an estimate of RR. The techniques used for this component act in either the time or frequency domain. Time-domain techniques involve detecting individual breaths in a respiratory signal. The RR is then calculated as the mean breath duration. Frequency domain techniques mostly involve calculating a frequency spectrum, from which the RR is identified as the frequency corresponding to the maximum spectral power in the range of plausible respiratory frequencies. The techniques are listed in Table 4.7. These techniques are used to estimate RR from a window of a respiratory signal, such as a 32 s window. The RR estimation component may be the last in a RR algorithm. However, two

TABLE 4.7: Techniques for RR estimation. Techniques used in this thesis are indicated by *

Abbr.	Technique
	Frequency-based, E_{F1} to E_{F7}
E_{F1}*	Fast Fourier Transform spectral analysis [80].
E_{F2}*	Auto-regressive spectral analysis [259] using model order 8 [237].
E_{F3}*	Auto-regressive spectral analysis using the median spectrum for model orders 2-20 [295].
E_{F4}*	Auto-regressive all-pole modelling (order 8), with the highest magnitude pole selected as the respiratory pole [177].
E_{F5}*	Auto-regressive all-pole modelling (order 8), with the lowest frequency pole selected as the respiratory pole [108].
E_{F6}*	Find periodicity using the autocorrelation function [109].
E_{F7}*	Spectral analysis using the Welch periodogram [199].
E_{F8}	Spectral analysis using the short-time Fourier transform [253].
E_{F9}	Use Gaussian process regression to estimate periodicity [242].
	Time domain breath detection, E_{T1} to E_{T5}
E_{T1}*	Breath detection by peak detection [280].
E_{T2}*	Breath detection by positive gradient zero-crossing detection [186].
E_{T3}*	Breath detection by combined peak and trough detection [281]: elimination of peaks less than the mean, and troughs greater than the mean; elimination of peaks (and troughs) within 0.5s of the previous peak (or trough); elimination of peaks (and troughs) which are immediately followed by a peak (or trough).
E_{T4}*	Breath detection using *Count-orig* [109]: detrend; detect peaks and troughs; define a threshold as 0.2 times the 75th percentile of peak values; ignore peaks with an amplitude below this threshold; identify valid breaths as consecutive peaks separated by only one trough with an amplitude less than zero.
E_{T5}*	Breath detection using *Count-adv* [109]: detrend; detect peaks and troughs; define a threshold as 0.3 times the 75th percentile of amplitude differences between consecutive extrema; eliminate the pair of extrema with the smallest amplitude difference if this is below the threshold; repeat until no more pairs can be eliminated; remaining peaks represent breaths.

further components can optionally be used to improve the robustness of RR estimates, and are now described.

TABLE 4.8: Techniques for fusion of respiratory rate estimates. Techniques used in this thesis are indicated by *

Abbr.	Technique
	Modulation, F_{M1} to F_{M4}
F_{M1}*	Smart Fusion [80]: RRs estimated from BW, AM and FM respiratory signals ($X_{B1,2,3}$) are quality assessed. If their standard deviation is ≤ 4 then RR is estimated as the mean, otherwise no RR is output.
F_{M2}*	Spectral peak-conditioned averaging [199]: Frequency spectra calculated from BW, AM and FM respiratory signals ($X_{B1,2,3}$) using the Welch periodogram (F_{T7}) are fused to give a mean spectrum. Only those spectra for which a certain proportion of spectral power is contained within a frequency range centred on the frequency corresponding to the maximum spectral power are included (a modification of the reported method). RR is estimated as the frequency corresponding to the maximum power in the mean spectrum.
F_{M3}*	Pole Magnitude Criterion [237]: The respiratory pole is chosen as the highest magnitude pole obtained from auto-regressive spectral analysis of BW, AM and FM respiratory signals ($X_{B1,2,3}$).
F_{M4}*	Pole Ranking Criterion [238]: The pair of highest magnitude poles obtained from auto-regressive spectral analysis of BW, AM and FM respiratory signals ($X_{B1,2,3}$) with the greatest pole ranking criterion (PRC) is selected. $PRC = \overline{m_{i,j}}/d\theta_{i,j}{}^2$, where $m_{i,j} = (m_i + m_j)/2$ and $d\theta_{i,j} = \lvert\theta_i - \theta_j\rvert$, for $i, j = 1, \ldots, N$, where N is the number of poles calculated. θ and m are the pole angles and magnitudes respectively. RR is estimated from the mean frequency corresponding to the selected pair.
	Temporal, F_{T1}
F_{T1}	Temporal smoothing [113]: estimated RRs, RR_{est}, are smoothed to give the final RR, RR_i, using $RR_i = 0.2RR_{est} + 0.8RR_{i-1}$.

4.4.3.4 Fusing respiratory rate estimates

Techniques for fusion of multiple RR estimates can be used to improve the robustness of a final RR estimate. They can be separated into two categories: modulation, or temporal fusion techniques. Modulation fusion techniques consist of fusing RR estimates derived from the same window of input signal using different modulations. For instance, simultaneous estimates derived from respiratory signals indicative of AM and FM have been fused to provide a single output. Temporal fusion techniques involve fusing RR estimates derived from a single respiratory signal, over different windows. For instance, when continuously monitoring RR during exercise testing previous RR estimates can be fused with the current RR estimate to improve robustness. The techniques which have been proposed for fusion of RR estimates are described in Table 4.8.

4.4.3.5 Quality assessment

Quality assessment is an optional component of RR algorithms, which can be performed at any stage of an algorithm. Quality assessment techniques used in RR algorithms fall into two categories: signal quality indices (SQIs) and RR estimate quality indices (RQIs). SQIs are used to identify segments of ECG or PPG data of low quality. Low quality segments are typically rejected based on the assumption that RR estimates derived from them are likely to be inaccurate [80]. RQIs are used to assess the expected accuracy of estimated RRs and reject any estimates which are expected to be inaccurate.

Several SQI techniques have been used in RR algorithms to avoid estimating RRs from low quality input data. SQIs based on template matching, described in Section 3.3, have been commonly used in RR algorithms. These involve construction of a template of average beat morphology, and calculation of the correlation between individual beats and the template. A signal segment is deemed to be of high or low quality by comparing the average correlation coefficient over that segment to an empirically-determined threshold. Another common approach is to use Hjorth parameters as SQIs [241, 296]. Hjorth parameters quantify the quality of oscillatory signals [297]. The highest quality is achieved by a signal consisting of a single oscillatory component, and the lowest quality by a signal which has a uniform frequency spectrum (*i.e.* white noise). If the Hjorth mobility, an estimate of a signal's central frequency, is not a plausible HR, then the signal is deemed to be of low quality. If the Hjorth complexity, an estimate of a signal's half-bandwidth, is small this indicates a strongly oscillatory signal. Conversely, if it is greater than an empirical threshold (e.g. 8 Hz), this indicates a non-oscillatory, and therefore low quality, signal [298]. An alternative approach is to search for any beat-by-beat characteristics which are indicative of low quality input signals. Beat-to-beat intervals [80, 280, 281], pulse amplitudes [80] and clipped pulses [80] have been used to identify segments of low quality input signal. SQIs are commonly used to identify and reject low quality input data in critical care [114], so many techniques are borrowed from this literature.

Several quality assessment techniques have been developed specifically for RR algorithms which assess either the extracted respiratory signals or the RR estimates. Similarly to ECG and PPG signal quality assessment, Hjorth parameters [111] and breath-to-breath intervals [280] have been used to assess the quality of respiratory signals. In the frequency domain, it has been proposed that if the ratio of the magnitude of the respiratory peak to that of the next highest peak is below a cut-off threshold, then the respiratory signal should be identified as

TABLE 4.9: Characteristics of RR algorithms assessed in the literature. Definition: PTT - pulse transit time

Characteristic	Category	No. publications (%)
Input Signal	ECG	71 (50.7)
	PPG	79 (56.4)
	ECG and PPG	1 (0.7)
	PTT	6 (4.3)
Window duration (s)	< 30	3 (2.1)
	30-59	30 (21.4)
	60-89	33 (23.6)
	≥ 90	7 (5.0)
	unknown	67 (47.9)
No. algorithms assessed	1	65 (46.4)
	2-5	58 (41.4)
	6-10	12 (8.6)
	11-15	3 (2.1)
	≥ 16	2 (1.4)

low quality [191]. The quality of RR estimates has been assessed by estimating three estimates simultaneously using the three respiratory modulations [80]. If the range of the estimates is small, then they are deemed to be of high quality.

4.4.4 Respiratory rate algorithm assessment methodologies

The characteristics of the RR algorithms assessed in the literature are provided in Table 4.9. Research into RR algorithms has been divided between the use of ECG and PPG signals as the input signal (used in 50.7 % and 56.4 % of publications respectively). Very few publications have assessed algorithms which use both the ECG and PPG (0.7 %), or the pulse transit time (PTT) as inputs (4.3 %). The duration of input signal used to estimate a single RR was most commonly between 30 and 90 s, although the duration was not reported in 46.4 % of publications. Very few publications conducted large scale comparisons of algorithms, shown by only 5 publications (3.6 %) containing comparisons of more than ten algorithms, and only 2 (1.4 %) comparing more than 15 algorithms. Approximately half of publications (46.4 %) assessed the performance of only one algorithm.

The characteristics of the datasets used to assess the performances of RR algorithms are provided in Table 4.10. The majority of publications used datasets containing data from young adults (63.6 %). In contrast only 30.0 % of publications used data from elderly adults. Most publications used data from healthy subjects (70.7 %). The most frequently used patient population was critically-ill patients (24.3 %), whereas only 5.7 % of publications used data from acutely-ill patients. Similarly, the data used in most publications were acquired during spontaneous breathing (77.1 %). Few publications assessed the performance of RR algorithms on ambulatory data (11.4 %). The duration of recordings used to assess RR algorithms varied greatly from $<$ 10 mins to several hours. Nearly a quarter (23.6 %) of studies did not report the duration of recordings used.

At least some of the data used in most assessments (74.3 %) is not publicly available. A total of 11 publicly available datasets have been used to assess the performance of RR algorithms, as detailed in Table 4.11. The publicly available datasets tend to be relatively small, with only three datasets containing data from more than 50 subjects. All of the datasets contain ECG signals, whereas only six out of eleven contain PPG signals. A range of respiratory signals were used to estimate reference RRs across these datasets. Data in five of the datasets, including the four largest datasets, were collected from critically-ill patients who may either be breathing unassisted or with the aid of a mechanical ventilator. Only three datasets contained data from sick patients outside of the critical care setting, all of which were recorded during sleep apnea assessments. Despite the public availability of 11 datasets with which to assess RR algorithms, only one publication has used more than 2 datasets [155]. Note that the VORTAL dataset has not been included here since it is described in Section 4.6.

The methods used to assess algorithm performance are summarised in Table 4.12. A range of equipment was used to acquire reference RRs, measurement of air flow or pressure, and impedance pneumography (ImP), being most common (each used in 22.1 % of publications). The type of reference RR equipment was not reported in 17.9 % of studies. A range of algorithms were used to obtain reference RRs from reference respiratory signals. However, often there was no assessment of the performance of these algorithms, making it difficult to know whether errors in RR estimates derived from the ECG and PPG were solely due to poor RR algorithm performance, or contributed to by inaccuracies in the algorithms used for obtaining reference RRs. Two notable exceptions were [109] and [10]. In [109] several algorithms for obtaining reference RRs were compared, and the time domain E_{T4} and E_{T5} methods were found to be *"the only serious candidates"*, with frequency domain spectral methods (E_{F1}, E_{F2}) and an

TABLE 4.10: Characteristics of the datasets used to assess RR algorithms in the literature.

Characteristic	Category	No. publications (%)
Age of subjects (years)	0 - 0.1: Neonate	3 (2.1)
	0.1 - 17: Paediatric	16 (11.4)
	18 - 39: Young adult	89 (63.6)
	40 - 69: Middle-aged adult	46 (32.9)
	$\geq$ 70: Elderly adult	42 (30.0)
Level of illness	Healthy	99 (70.7)
	Sick in community	18 (12.9)
	Acutely-ill	8 (5.7)
	Critically-ill	34 (24.3)
	Unknown	5 (3.6)
Breathing	Spontaneous	108 (77.1)
	Metronome	33 (23.6)
	Ventilated	26 (18.6)
	Unknown	15 (10.7)
Included ambulatory data?	Yes	16 (11.4)
	No	134 (95.7)
	Unknown	1 (0.7)
Recording time (mins)	1-10	50 (35.7)
	11-20	20 (14.3)
	21-59	9 (6.4)
	$\geq$ 60	28 (20.0)
	Unknown	33 (23.6)
Dataset publicly available?	Yes	60 (42.9)
	No	104 (74.3)
Number of datasets used	1	122 (87.1)
	2	17 (12.1)
	3	1 (0.7)

TABLE 4.11: Publicly available datasets suitable for the assessment of RR algorithms. Definitions: Age - paediatric (paed), young adult (young), elderly adult (elderly); Inputs - seismocardiogram (SCG); Resp Sigs - capnometry (CO_2), piezoresistive thoracic band (piezo), oral or nasal flow (flow), inductance plethysmography (InP), impedance pneumography (ImP); Breathing - spontaneous (spont), ventilated (vent); unknown (unk)

Dataset	Ref	No subjs	Age	Inputs	Resp Sigs	Breathing	Level of Illness
MIMIC-II	[299]	23,180	paed, young, elderly	ECG, PPG	ImP	spont, vent	critical
MGH/MF	[300]	250	paed, young, elderly	ECG	ImP, CO_2	spont, vent	critical
MIMIC	[301]	72	unk	ECG, PPG	ImP	spont, vent	critical
CapnoBase	[80]	42	paed, young, elderly	ECG, PPG	CO_2	spont, vent	critical
Fantasia	[302]	40	young, elderly	ECG	unk	spont	healthy
UCD Sleep Apnea	[303]	25	young, elderly	ECG, EMG	flow	spont	healthy, apnea
CEBS	[304]	20	young	ECG, SCG	piezo	spont	healthy
MIT-BIH Polysomnographic	[305]	18	young	ECG	flow	spont, vent	healthy, apnea
Stress Recognition in Automobile Drivers	[306]	17	unk	ECG, EMG	unk	spont	healthy
Apnea-ECG	[307]	8	young, elderly	ECG, PPG	InP, flow	spont	healthy, apnea
Portland State	[308]	1	paed	ECG, PPG	unk	unk	critical

TABLE 4.12: Methods used to assess RR algorithm performance in the literature.

Characteristic	Category	No. publications (%)
Reference RR equipment	air flow or pressure	31 (22.1)
	ImP	31 (22.1)
	InP	11 (7.9)
	Piezoelectric	7 (5.0)
	Strain gauge	10 (7.1)
	Metronome	7 (5.0)
	Other	14 (10.0)
	None	3 (2.1)
	Unknown	25 (17.9)
Statistics	Error statistic	71 (50.7)
	Breath detection statistic	25 (17.9)
	Bias and LOA	18 (12.9)
	Correlation	8 (5.7)
	Coverage Probability	2 (1.4)
	None	28 (20.0)
	Unknown	3 (2.1)
Conflict of interest?	Yes	1 (0.7)
	No	139 (99.3)

autocorrelation method (E_{F6}) performing poorly. In [10] (which arose from this review) a time domain breath detection algorithm was used, and its performance was quantified by comparing the reference RRs provided by the algorithm to those obtained through manual annotation of a subset of the data.

Similarly, a wide range of statistics were used to assess RR algorithm performance. The most common group of statistics were based on the errors in estimated RRs (used in 50.7 % of publications), including the mean (absolute) error, root mean squared (normalised) error, and the percentage error. Statistics indicating the reliability of breath detection were used in 17.9 % of publications. These included statistics such as sensitivity, specificity, false negative and false positive rates. The limits of agreement method [309], consisting of the systematic bias and limits of agreement within which 95% of errors are expected to lie, was used less often (12.9 %). Only one publication reported a conflict of interest.

4.5 Discussion

This review has presented a comprehensive synthesis of the literature on algorithms to estimate RR from the ECG and PPG. A total of 140 publications were included in the review following a systematic search of four online databases. This demonstrates both the importance of the topic to the research community, and the potential utility of a comprehensive review to future researchers. This review builds on previous reviews of algorithms for use with the ECG [79] and PPG [280, 281, 310]. The results of the review are now discussed.

4.5.1 Search and screening

The methodology used to conduct this review differed from previous reviews of the topic, which consisted of manual search strategies. In contrast, this review used a combination of manual and systematic search strategies, ensuring a highly comprehensive search. The systematic search strategy was designed based on the results of an initial manual search, allowing the search terms to be optimised for identification of relevant literature. This resulted in a relatively high PPV for the identified publications of 25.3 %, indicating the proportion of publications identified based on their title which were included. In addition to the manual search, four electronic databases were searched, providing a comprehensive survey of the literature. The most comprehensive search database was Scopus, which identified 91 included publications, including 65.0 % of the publications which were included. IEEE Xplore had a high PPV of 48.8 %, indicating its usefulness for quick, initial searches. However, it only identified 42.1 % of the included publications. Google Scholar had low sensitivity and PPV values (25.7 and 10.7 % respectively), and only provided 7 unique publications. The manual search strategy was highly important since 31 (22.1 %) of the included publications were only identified by the manual search and not any of the electronic databases. It was, however, by far the most time-consuming search. The search strategy used was not only beneficial because of its comprehensiveness, but has the additional advantage that it can be easily updated in the future.

4.5.2 Characteristics of included publications

This review demonstrated that research into the estimation of RR from the ECG and PPG is both a recently developed and rapidly growing field. The earliest included publication was

published in 1991. Since then the rate of publication has increased greatly, and continues to grow. Currently there are approximately 16 new publications describing RR algorithms per year. This highlights the need for frequent updates of this review, which are facilitated by the systematic search strategy.

4.5.3 Respiratory rate algorithms

This review presented a generalised structure of RR algorithms which can be used to facilitate in-depth assessments of algorithm performance. Algorithms were found to consist of up to five stages, as summarised in Figure 4.4. Two of the stages are compulsory: extraction of respiratory signal(s), and estimation of RR(s). The remaining three stages are optional. An algorithm can be constructed by selecting an interchangeable technique for use at each stage of the algorithm. This framework is a novel approach to decomposing RR algorithms into their constituent parts, which allows the multitude of different algorithms described in the literature (over 100) to be described concisely. It will also facilitate assessment of the influence of the technique used at each stage on algorithm performance. Some algorithms contain modifications, or additions, to the orthodox techniques listed for use at each stage. For instance, techniques for estimation of RR based on autoregressive modelling often require specification of the model order. Several approaches have been proposed to select an appropriate model order, including: (i) the use of a pre-specified model order [108, 237, 280]; (ii) analysis of the respiratory signal using Rissanen's minimum description length method [110, 311, 312] or the Akiake criteria [111]; or (iii) the optimal parameter search (OPS) criterion [203–205, 271]. Therefore, the techniques listed in the generalised structure do not capture all of the potential subtleties of algorithms. However, this structure has the great benefit of providing a manageable way to implement and assess the many algorithms proposed in the literature.

The extraction of respiratory signals was performed using two approaches: feature- and filter-based approaches. Feature-based techniques extract one measurement of the influence of respiration on the input signal for each heart beat. In contrast, filter-based techniques extract a continuous signal. The benefit of the feature-based approach is that a specific influence can be measured. For instance, the change in QRS amplitude due to movement of the ECG electrodes with respect to the heart's electrical axis can be measured directly, without any confounding from P- or T-waves. In contrast, filter-based techniques make no distinction between the different components of the ECG or PPG signals, and therefore do not measure a specific influence.

However, the feature-based approach has the disadvantage of only taking a measurement of the influence of respiration once per heart beat. If the HR is less than twice the RR, then this may inhibit RR estimation by the Nyquist-Shannon sampling theorem [313]. Conversely, the filter-based approach provides a continuous signal, potentially allowing estimation of RRs when the HR is less than twice the RR. It is not yet clear which approach provides superior performance, or whether the optimal choice of approach varies between patient populations. Furthermore, it is not clear which technique for extracting a respiratory signal provides the highest quality signal (*i.e.* that which correlates most closely with a reference respiratory signal), and whether this also varies between patient populations.

The fusion of multiple respiratory signals to provide one, more robust signal, has received relatively little attention in the literature. Time domain methods have received little consideration since their conception in 2003 [186]. Frequency domain methods have been used in a few publications, and often have the benefit of incorporating a quality assessment step, for instance only including respiratory signals whose frequency spectra meet certain criteria.

Estimation of RRs was performed using techniques operating in either the time or the frequency domain. Time domain techniques consist of the identification of individual breaths. These techniques have the benefit of being able to estimate RRs when the respiratory pattern is not stationary. However, they are susceptible to erroneous breath detection due to abnormal morphology in a respiratory signal. On the other hand, frequency domain techniques require the respiratory pattern to be quasi-stationary, but are less susceptible to transient noise on the respiratory signal. Similarly to feature- and filter-based approaches to extracting a respiratory signal, it is not yet clear whether either approach for estimation of RRs provides superior performance.

The fusion of multiple RR estimates has been proposed as a method to increase the robustness of RR estimates [80, 237, 238]. However, these techniques have not been tested widely across different datasets, so it is not clear whether this optional component consistently improves algorithm performance. The techniques fall into two categories: modulation and temporal fusion. Modulation fusion is conducted by fusing individual estimates derived from multiple respiratory signals, whereas temporal fusion consists of fusing sequential estimates from a single respiratory signal. Therefore, temporal fusion may not be suitable for use with wearable sensors since there may be gaps between consecutive segments of high quality data.

A range of quality assessment techniques have been proposed. However, no individual technique has been commonly used across several publications, highlighting the experimental nature of these techniques. Recent work has compared the performances of several techniques for assessment of respiratory signal quality, although this work is still ongoing [154, 155].

4.5.4 Respiratory rate algorithm assessment methodologies

The 140 included publications used a wide range of methodologies for assessment of RR algorithms. Research has been split almost equally between using the ECG and PPG as input signals. This demonstrates the potential clinical utility of using each input signal, as well as the need to compare algorithm performances when using each input signal. The duration of input signals used to estimate RR ranged between 30 and 90 s in most assessments. This reflects a compromise between, on the one hand, reducing the time required for assessments and ensuring that the true RR remains stable during the measurement period, and on the other hand, increasing the measurement duration to obtain more precise estimates.

The focus of RR algorithm assessments appears to have been on the assessment of novel algorithms, rather than comparisons of performances of existing algorithms. This is demonstrated by approximately half of the publications assessing the performance of only one algorithm. In contrast, only two publications (one of which arose from this work) compared the performance of more than 15 algorithms. Consequently, it is difficult to determine from the literature which algorithms perform best.

The datasets used to assess RR algorithms were often acquired from subjects who are not representative of acutely-ill, ambulatory patients who would be monitored with wearable sensors. Assessments were often conducted using data from young, healthy subjects. A smaller proportion used data from ventilated patients or subjects breathing in time with a metronome, whose respiratory mechanics could not be presumed to be similar to those of spontaneously breathing patients. In contrast many hospital patients are elderly, and therefore have stiffer arteries, more frequent ectopic beats, and co-morbidities which may affect the performance of RR algorithms. Only a few publications (8, 5.7 %) used data from acutely-ill patients. This is likely to be because it is difficult to acquire reference RR measurements in acutely-ill, ambulatory patients, and also because none of the publicly available datasets contain data from acutely-ill patients outside of critical care.

4.5.5 Future Research Directions

This review identified three key limitations of previous assessments of RR algorithms which should be addressed in future research. Firstly, prior to this review there were no comprehensive comparisons of RR algorithms. The performances of 30 algorithms were compared in one publication [278], whilst the only other publication to have compared more than 15 RR algorithms arose from this review [10]. This demonstrated the urgent need for a comprehensive assessment of RR algorithms to determine whether any algorithm performs sufficiently well for use in wearable sensors. Secondly, most of the assessments used data from populations which are not necessarily representative of acutely-ill patients. Therefore, it is not known whether RR algorithms could be expected to perform at the same level when used for monitoring hospital patients using wearable sensors. Thirdly, it is unclear whether the reference RRs used in many publications were reliably obtained, since the algorithms used to obtain them were often not assessed. This may have resulted in the observed performances of RR algorithms being worse than their true performances due to inaccuracies in reference RRs. The impacts of these three limitations on future research are now discussed.

Comprehensive assessments of algorithms are currently hindered by the lack of publicly available algorithm implementations. The publications identified were written by 101 first authors, indicative of the wide range of skills and knowledge used to design the algorithms. Therefore, a substantial amount of effort is required to implement the algorithms for assessment. An alternative approach which would not require implementation of the different algorithms would be to perform a meta-analysis of existing studies. However, several factors make this difficult including: poor reporting of study methodology; the use of different statistical methods to assess algorithms; the use of data from several populations ranging from healthy subjects to critically-ill patients; the potential for different algorithm implementations between studies; and the potential range of accuracies of reference RRs. Therefore, a publicly available toolbox of algorithms is required to facilitate comprehensive assessments.

Assessments of the utility of algorithms for use with wearable sensors are hindered by a lack of datasets acquired from acutely-ill patients. None of the 11 publicly available databases contain data from acutely-ill patients outside of the critical care setting. Therefore, additional datasets are required to assess the potential performance of algorithms in this setting.

It is essential that the reference RRs used to evaluate RR algorithms are reliable. However, there has been little research into the techniques used for obtaining reference RRs [237]. Consequently, it is not clear how accurate reference RRs are, and also whether the technique used to obtain reference RRs biases the performance metrics towards particular methods for estimating RR from the ECG and PPG. These are important considerations since inaccuracies in reference RRs will result in the performances of RR algorithms appearing to be poorer than they truly are. Several methods have been used previously to obtain reference RRs, although none are ideal. These include: manual annotation of individual breaths [80]; using a time domain breath detection algorithm to identify individual breaths [109, 204]; using two methods for estimating RR, and only outputting a reference RR at times when the two methods agree within a specified tolerance (*e.g.* two frequency domain techniques [154], or a time domain and a frequency domain technique [4, 237]). However, none of these methods is without its limitations. Manual annotations are time-consuming and not easily reproducible [109]. The use of a time domain algorithm may bias the analysis towards RR algorithms which use a time domain RR estimation technique. The use of two methods to estimate RR from the reference signal may result in excessive exclusion of windows, since those windows on which only one of the two methods provides an accurate reference RR will be excluded. It may not be possible to develop a perfect technique for obtaining reference RRs which is free from any of these limitations. However, further research should be based on an informed choice of technique, allowing its limitations and possible influence on the results to be appreciated.

Despite the extensive research into RR algorithms, none have yet been widely adopted into clinical practice. However, clinical devices are now available which incorporate RR algorithms. Covidien^{TM} have developed an RR algorithm for use with the PPG which can be incorporated into their patient monitors. This is likely based on the algorithm they have previously reported and assessed in healthy volunteers and hospitalised patients [147, 148]. The device has been independently assessed [260], although it is not yet clear whether it is suitable for widespread clinical use. In addition, Philips Healthcare (Andover, MA, USA) have conducted studies investigating the estimation of RR from the ECG [151, 153, 176, 183–185]. The availability of clinical devices containing an RR algorithm makes it substantially easier to assess RR algorithms prospectively in a range of clinical settings, although such research is limited to the algorithm implemented in the device.

4.6 RRest: Respiratory Rate Estimation Resources

This section presents the development of three resources to address the three limitations of RR algorithm assessments described above: a lack of datasets from acutely-ill patients; the need to ensure that reference RRs are reliable; and, a lack of publicly available algorithm implementations. These resources are designed for two purposes. Firstly, they are used in the remainder of this thesis to assess RR algorithms for retrospective use with the wearable sensor data in the LISTEN Dataset. Secondly, they are being made publicly available, enabling researchers to evaluate novel algorithms, and to perform their own assessments of RR algorithms.

4.6.1 Datasets for Assessment of RR Algorithms

The range of datasets publicly available for assessment of RR algorithms is not sufficient to determine whether RR algorithms perform well enough for use with wearable sensors. The publicly available datasets which have previously been used to assess RR algorithms (listed in Table 4.11) are not ideal for two reasons. Firstly, the available datasets were recorded from three patient populations, none of which is entirely representative of acutely-ill patients who would wear wearable sensors. These are: (i) critically-ill patients; (ii) healthy subjects; and, (iii) subjects being assessed for chronic obstructive sleep apnoea. Critically-ill patients often require organ support which may impact the respiratory, cardiovascular and autonomic nervous systems, artificially altering their physiology. For instance, patients may require mechanical ventilation which provides positive pressure during inhalation (rather than the usual negative pressure provided by thoracic expansion). They may also require positive or negative inotropic agents which increase or decrease the strength of heart muscle contraction respectively. Critically-ill patients may also be subject to pathologies such as sepsis which reduce autonomic nervous system functionality [314]. Conversely, healthy subjects may have improved organ function when compared to acutely-ill patients. Finally, sleep apnoea markedly changes respiratory dynamics, making it unrepresentative of acutely-ill patients. Secondly, only in vivo datasets are available, making it difficult to ensure that algorithms have been reasonably implemented, since when using in vivo datasets it is difficult to determine whether poor algorithm performance is due to a lack of respiratory modulation of the ECG and PPG signals, or due to poor algorithm implementation. Therefore, there is a need for additional datasets for initial verification of RR

algorithm implementations, and subsequent assessment of the performances of RR algorithms in subjects who are similar to the target population of acutely-ill patients.

In the remainder of this section six novel datasets are described which have been assembled to facilitate the assessment of RR algorithms for use in wearable sensors. These have been designed to allow the influences of individual physiological and technical factors on RR algorithms to be assessed. The datasets are summarised in Table 4.13.

RRest-synth: a dataset of idealised synthetic signals

Currently it is not possible to verify algorithm implementations on an idealised dataset. Rather, algorithms are implemented and tested on real-world datasets. This approach makes it difficult to perform robust comparisons of algorithms, since it is difficult to determine whether poor performances of algorithms are due to intrinsic flaws in the algorithmic techniques, or simply due to flawed implementations of the algorithms. The RRest-synth Dataset was designed to overcome these difficulties. The dataset contains simulated ECG and PPG signals covering a range of heart rates (HRs) and RRs. Three signals are provided for each combination of HR and RR, each exhibiting either baseline wander (BW), amplitude modulation (AM), or frequency modulation (FM). The exemplary signals shown in Figure 4.1 are taken from this dataset.

The signals were simulated based on mathematical descriptions of ECG and PPG signals in the presence of the three respiratory modulations: BW, AM and FM. This builds on previous work by Pallás-Areny *et al.*, who described a mathematical formulation for AM of the ECG [315]. They assumed that if the ECG (here extended to the PPG), denoted $x(t)$, is assumed to be periodic, then it can be described as

$$x(t) = \sum_n V_n \cos(n\omega_c t + \Phi_n) \quad , \tag{4.1}$$

where the periodic signal consists of n sinusoidal components of amplitude V_n, ω_c is the constant cardiac angular frequency (frequency, $f = 2\pi\omega_c$), and Φ_n is a phase shift [315]. This describes the ECG or PPG in the absence of any respiratory modulation.

The manifestation of BW on the unmodulated signal can be modelled by assuming that the respiratory modulation is sinusoidal, at constant respiratory angular frequency ω_r with amplitude A_r, and that it obeys the superposition principle. Therefore, an ECG or PPG signal influenced

TABLE 4.13: Novel datasets for assessment of RR algorithms. Definitions: Age - young adult (young), elderly adult (elderly); Resp Sigs - impedance pneumography (ImP), oral or nasal pressure (press); Breathing - spontaneous (spont), controlled (cont)

Dataset	Ref	No subjs	Age	Inputs	Resp Sigs	Breathing	Level of Illness
RRest-synth	[10]	n/a	n/a	ECG, PPG	ref RRs	simulated	healthy
RRest-ideal	[10]	39	young	ECG, PPG	ImP, press	spont	healthy
RRest-healthy	[7]	57	young, elderly	ECG, PPG	ImP, press	spont	healthy
RRest-acute	-	108	young, elderly	ECG, PPG	ImP	spont	acute
RRest-afib	-	15	young, elderly	ECG, PPG	ImP	spont	acute
RRest-vent	-	59	young, elderly	ECG, PPG	ImP	cont, spont	critical, acute

by BW can be expressed as

$$x(t) = \sum_n V_n \cos(n\omega_c t + \Phi_n) + A_r \cos(\omega_r t + \Phi_r) \quad , \tag{4.2}$$

where Φ_r is the respiratory phase shift.

The manifestation of AM can be modelled through time domain multiplication of the amplitude of the unmodulated signal by a respiratory modulation (the latter being always positive). The respiratory modulation is characterised by an oscillation of maximum amplitude modulation ΔV and at angular frequency ω_r. Therefore, an ECG or PPG signal influenced by AM can be expressed as

$$x(t) = \sum_n V_n \cos(n\omega_c t + \Phi_n) \cdot (1 + \Delta V \cos \omega_r t) \quad . \tag{4.3}$$

The manifestation of FM can be modelled by modulating the instantaneous cardiac angular frequency, $\omega_{ci}(t)$, at the respiratory frequency, giving

$$\omega_{ci}(t) = \omega_c + \Delta\alpha \sin(w_r t) \quad , \tag{4.4}$$

where $\Delta\alpha$ is the maximum angular displacement. The instantaneous phase angle of the FM signal, $\theta_{ci}(t)$, is found by integrating $\omega_{ci}(t)$ over time, as

$$\theta_{ci}(t) = \omega_c t - \frac{\Delta\alpha}{w_r} \cos(w_r t) \quad . \tag{4.5}$$

Therefore, an ECG or PPG signal influenced by FM can be expressed as

$$x(t) = \sum_n V_n \cos(n[\omega_c t - \frac{\Delta\alpha}{w_r} \cos(w_r t)] + \Phi_n) \quad . \tag{4.6}$$

The dataset was generated as follows. An exemplary beat for the ECG and PPG signal was acquired from a young subject. The time values of these beats were interpolated so that each beat lasted 1 s. These were then repeated 210 times, giving a simulated signal lasting 210 s. This signal was then modulated according to the descriptions of the three respiratory modulations listed above, producing three separate signals. This process was repeated for a range of HRs (30 - 200 beats per minute) and RRs (4 - 60 breaths per minute). When the HR was varied, the RR was fixed at 20 bpm. When the RR was varied, the HR was fixed at 80 bpm. This provided a total of 192 signals, each sampled at 500 Hz.

The RRest-synth Dataset serves two purposes. Firstly, it allows one to determine whether an RR algorithm has been implemented reasonably (i.e. whether it estimates RR accurately in ideal simulated signals). Secondly, it allows one to assess the limitations of RR algorithms, such as whether they perform accurately in the presence of different types of respiratory modulation, and whether their performance is dependent on the underlying HR or RR.

RRest-ideal: a dataset of young healthy subjects before, during and after exercise

It is useful to assess the performance of RR algorithms on data from healthy subjects as well as patients, to determine their best possible performance in the absence of pathophysiology. Despite five publicly available datasets containing data from healthy subjects, none of these are ideal for this purpose. Only one of these five contains PPG signals, which were recorded from eight subjects. Therefore, a novel dataset acquired from healthy subjects is required to assess the performance of RR algorithms on both ECG and PPG signals.

The RRest-ideal Dataset contains data acquired from 39 young healthy subjects aged 18 - 39 years old as part of the VORTAL study (National Clinical Trial 01472133) [10]. Ethical approval for the study was obtained from the London Westminster Research Ethics Committee (11/LO/1667). Data were acquired for approximately 10 mins at rest in the supine position, followed by 2 mins walking on a treadmill, followed by several minutes running (until the subject's HR reached 85 % of their age-predicted maximum [316]), followed by a further 10 mins at rest whilst recovering from the exercise.

Therefore, the RRest-ideal Dataset is useful for two purposes. Firstly, it facilitates assessment of the performance of algorithms in ideal conditions. Secondly, it can be used to assess the impacts of walking and running on RR algorithm performances.

RRest-healthy: a dataset of young and elderly healthy subjects at rest

Several technical and physiological factors may influence the performance of RR algorithms, as described in Chapter 5. These include technical factors such as the use of high fidelity recording equipment (as opposed to clinical monitors), signal sampling frequency, the use of the ECG or PPG as the input signal, and the location of PPG measurement (*e.g.* ear or finger); and physiological factors such as age, gender, RR and HR. The technical factors may influence the

design of wearable sensors, whilst the physiological factors may influence the choice of clinical settings in which RR algorithms are used.

The RRest-healthy Dataset is designed to facilitate investigation of the influence of each of these factors on the performances of RR algorithms. It contains data acquired from 41 young healthy subjects aged 18 - 39 years old, and 16 elderly healthy subjects aged over 70 years old, as part of the VORTAL study [7]. A 10 minute recording was acquired from each subject whilst laid supine. Subjects who had co-morbidities or were receiving medications that might significantly affect the functioning of the cardiac, respiratory and autonomic nervous systems were excluded.

Data were acquired simultaneously using a range of sensors. High fidelity laboratory (lab) equipment was used to acquire lead II ECG, finger PPG, ear PPG, and oral-nasal pressure signals. The lab equipment consisted of a 1902 amplifier, a Power 1401 analogue-to-digital converter and Spike2 v.7.09 acquisition software (all Cambridge Electronic Design, Cambridge, UK). Finger and ear PPGs were transduced using MLT1020FC and MLT1060EC infra-red reflection plethysmographs respectively (AD Instruments, CO Springs, New Zealand). Oral-nasal pressure was transduced using an Ultima Dual Airflow differential pressure transducer (Braebon Medical Corporation, Kantata, ON, Canada) connected to a P1300 Pro-Flow oral-nasal cannula (Philips Respironics, Murrysville, PA, USA). Signals were sampled at 500 Hz. In addition, clinical equipment was used to simultaneously acquire Lead II ECG and finger PPG signals. The signals were monitored using an IntelliVue MP30 clinical monitor (Philips Medical Systems, Boeblingen, Germany) and captured using ixTrend acquisition software (v.2.0.0 Express, Ixellence GmbH, Wildau, Germany) at 500 Hz and 125 Hz, respectively.

The RRest-healthy Dataset is used in the investigation presented in Chapter 5, where additional details such as demographic characteristics are provided.

RRest-acute: a dataset of acutely-ill patients recovering from cardiac surgery

The RRest-acute Dataset is designed to facilitate assessment of the performances of RR algorithms in acutely-ill patients. Approximately 20 mins of data were acquired from each of 108 patients recovering from major cardiac surgery as part of the LISTEN Trial (National Clinical Trial 01549717). The data were acquired during the last eight hours of patients' stays in critical care prior to being discharged to an ambulatory ward where most were monitored using wearable sensors. Therefore, the physiology of the patients can be expected to be similar to that

which would be encountered when monitoring acutely-ill patients using wearable sensors. To be eligible, patients had to be in sinus rhythm, and not paced, excluding patients with arrhythmias or receiving external pacing.

RRest-afib: a dataset containing recordings during sinus rhythm and atrial fibrillation

The RRest-afib Dataset is designed for assessment of the influence of atrial fibrillation (AFib) on RR algorithms. The data, consisting of paired 20 min recordings of periods of AFib and sinus rhythm, was extracted from the LISTEN Dataset. All patients in the dataset were assessed for eligibility using heart rhythm labels documented in the routine clinical records. To be eligible, patients had to be in AFib for a sustained period, and in sinus rhythm for a separate sustained period, during the 24 hours of their critical care stay immediately prior to discharge to an ambulatory ward. A total of 20 patients were eligible according to the heart rhythm labels. Of these, five patients were removed because the labels were found to be inaccurate upon inspection. Therefore, RRest-afib contains 30 recordings, consisting of a recording during AFib and a recording during sinus rhythm, from 15 patients.

The reliability of the data labels was also assessed through analysis of the ECG R-R intervals in the dataset. This was achieved using a previously reported method for detecting AFib from R-R intervals [317]. This method quantifies the variability of R-R intervals by plotting the durations of the R-R intervals in each 128 s against the difference between consecutive R-R intervals. The plot is divided into a 25 ms resolution grid, and the number of non-empty cells (NECs, 25 $\times$ 25 ms cells) is calculated. Higher numbers of NECs demonstrate greater levels of variability, associated with AFib. The results, shown in Figure 4.7, show a marked difference in the variabilities of R-R intervals between the two groups.

RRest-vent: a dataset containing recordings shortly before and after extubation

The final dataset, RRest-vent, is designed to facilitate assessment of the impact of mechanical ventilation on RR algorithms. It consists of data acquired from 59 LISTEN Trial patients at three points during their recovery from surgery: shortly before disconnection from a ventilator; shortly after disconnection from a ventilator; and, shortly before discharge from critical care to the ambulatory ward. At each time point approximately 20 mins of data were acquired.

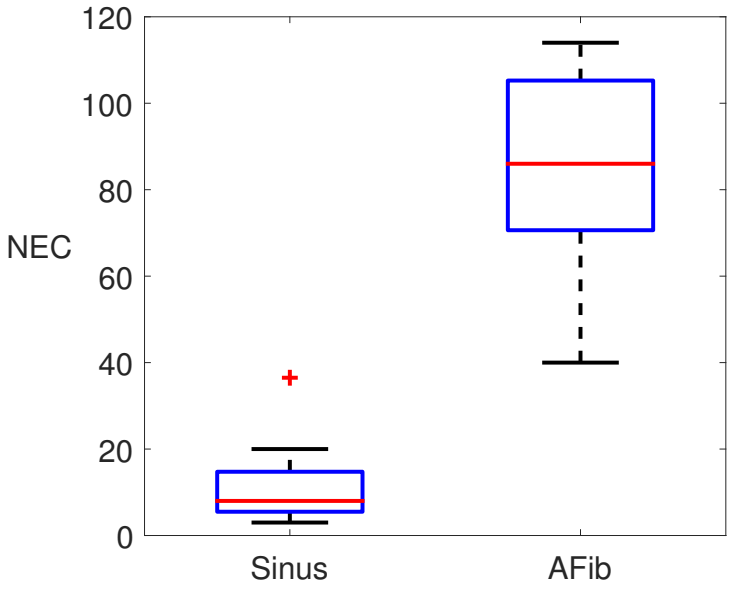

FIGURE 4.7: RRest-AFib Dataset verification: This shows the difference between the variabilities of the R-R intervals in the *sinus rhythm* recordings and *atrial fibrillation (AFib)* recordings. The distributions (median, lower and upper quartiles) of the NECs (a measure of variability of R-R intervals) are shown for each set of recordings. As expected, the AFib recordings exhibit a much greater level of variability than the sinus recordings.

This dataset will not be used to help directly with the assessment of algorithms for use with wearable sensors, since acutely-ill patients who may wear wearable sensors are not mechanically ventilated. Instead, assessment of the impact of mechanical ventilation on RR algorithms will contribute towards an understanding of whether the results of previous studies of RR algorithms in mechanically ventilated patients are representative of those which would be obtained during unassisted breathing.

4.6.2 A Signal Quality Index for the Impedance Pneumography Signal

One finding of the review was the need for further research into methods for obtaining reference RRs, as detailed in Section 4.5.5. All publicly available datasets for assessment of RR algorithms include a reference respiratory signal from which reference RRs can be obtained (Table 4.11), although only one (CapnoBase) contains reference RRs. Therefore, an algorithm is usually required to obtain reference RRs from the reference respiratory signal. Several algorithms have been proposed for this purpose as described in Section 4.5.5, although these each have significant limitations. An alternative approach would be to use a signal quality index (SQI) to discard segments of low quality reference respiratory signals, ensuring that only segments of high quality respiratory signals are used to obtain reference RRs. Indeed, this approach has been widely used to obtain high quality parameters from the ECG, PPG and ABP signals, as demonstrated in

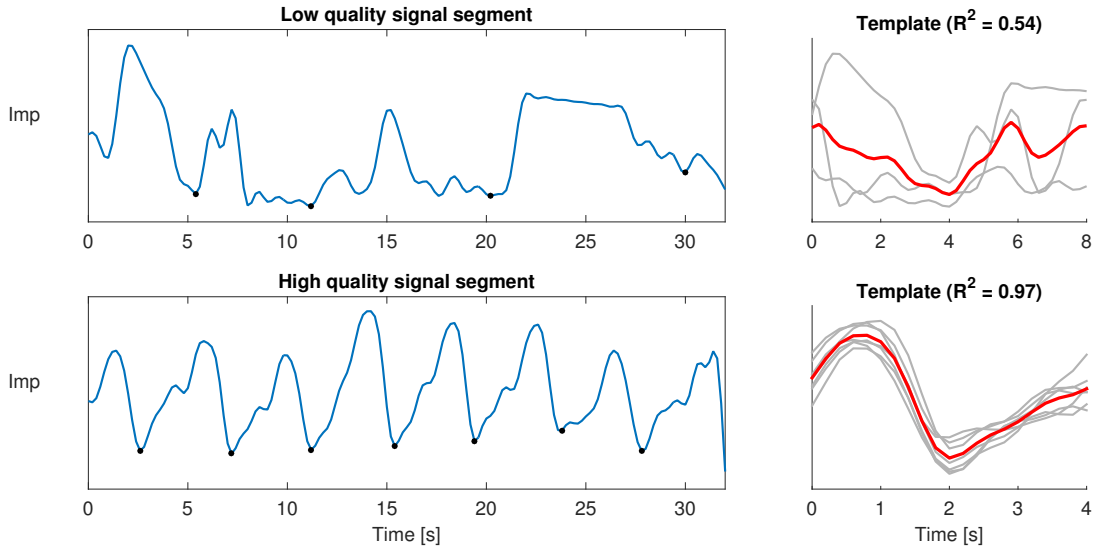

FIGURE 4.8: Impedance (ImP) signal quality assessment: A novel SQI algorithm was designed to assess the quality of ImP signal segments. On the left, two segments are shown with dots indicating detected breaths. On the right, the corresponding breath templates (red) and individual breaths (grey) are shown. The upper segment is of low quality, as indicated by a low correlation coefficient (R^2) of 0.54. The lower segment is of high quality, as indicated by a high R^2 of 0.97.

Section 4.4.3.5. However, SQIs have not yet been thoroughly developed for use with respiratory signals. A SQI has been previously proposed for use with the ImP signal [318]. However, when used for RR estimation it was reported to result in a relatively large mean error of 3.2 ± 4.6 bpm. Furthermore, it did not perform adequately in initial testing on this dataset. This section describes the development of a novel SQI for use with the ImP signal, and subsequent assessment of its performance.

The novel SQI was developed by adapting the approach proposed by Orphanidou *et al.* for cardiac signals in [124]. This approach consists of three steps. Firstly, individual respiratory cycles are detected in the signal (*i.e.* heart beats in the original implementation). Secondly, the timings of the individual cycles are assessed for physiological plausibility. For instance, in the original implementation all beat-to-beat intervals had to be less than 3 s in duration. Any signal segments with implausible cycle timings were deemed to be of low quality. Thirdly, template-matching was used to assess the similarity of the signal at each cycle as shown in Figure 4.8. If the correlation between the average cycle's morphology, and each individual cycle's morphology, was high enough then the signal segment was deemed to be of high quality.

A range of criteria for assessing the plausibility of breath-to-breath timings were trialled, resulting in the following algorithm:

1. **Pre-process the ImP signal:** The following steps were taken: low pass filter to remove frequency content above 1 Hz (60 bpm) using a Tukey window to avoid edge effects; downsample to 5 Hz; normalise to give a mean of 0 and standard deviation of 1; invert and remove any linear trend.

2. **Detect individual breaths:** The *Count-orig* method (E_{T4}) was used to detect breaths, as described in Table 4.7. This allows individual breaths to be labelled as either valid (high quality), or invalid (low quality).

3. **Assess plausibility of breath timings:** Three criteria were used: (i) the normalised standard deviation of breath durations had to be < 0.25 to permit only a small variation in the durations of detected breaths; (ii) the proportion of breath durations > 1.5, or < 0.5, times the median breath duration had to be < 15 % to prevent errors due to outlying breath durations; (iii) >60 % of the window duration had to be occupied by non-outlying breaths. Any window which did not satisfy these three criteria was deemed to be of low quality.

4. **Assess similarity of breath morphologies:** The morphologies of individual breaths had to be highly correlated, as demonstrated by a mean correlation coefficient between individual breath morphologies and a template breath of ≥ 0.75.

5. **Calculate RR:** If a window was deemed to be of high quality then the reference RR was calculated as the mean duration of the valid breaths in that window.

The performance of the SQI was assessed using the RRest-vent Dataset. This dataset was chosen to allow the performance to be assessed in three settings: ventilated in ICU, unassisted in ICU, and shortly before discharge from critical care to the ambulatory ward. The assessment was performed in two steps. Firstly, the ability of the SQI to discriminate between high and low quality signal segments was assessed. To do so, ImP signal segments were manually labelled as high or low quality, where a high quality was only given if the annotator was confident that they could accurately identify all the breaths in that segment. A total of 1801 signal segments were annotated (each of 32 s duration, totalling 16 hours of signals). These comprised approximately 10 mins of data from 34 subjects in each of the three clinical settings. Secondly, the accuracy and precision of reference RRs obtained using the SQI were assessed. To do so, individual breaths were annotated in those windows deemed to be of high quality by the SQI. These were used to calculate reference RRs to compare with those calculated by the SQI. In addition, a benchmark

method for identifying high quality respiratory signal segments, and obtaining reference RRs was implemented and assessed. The method chosen was the agreement SQI, used in [4, 237]. It consists of estimating RR from the respiratory signal using two independent methods, one in the time domain and one in the frequency domain. If the two estimates agree to within $\pm$ 2 bpm then the window is deemed to be of high quality, and the reference RR is calculated as the mean of the two estimates. The two methods used were E_{T4} and E_{F7}.

Results detailing the performance of both the novel ImP SQI, and the agreement SQI, are provided in Table 4.14. The results show the improved performance of the ImP SQI compared to the agreement SQI, across all performance metrics and all clinical settings. The key advantages of the ImP SQI were as follows. Firstly, it was able to provide better discrimination between high and low quality signal segments. It provided an increased sensitivity of 74.0 % compared to 48.6 % across all clinical settings, showing that more of the segments identified as high quality were truly of high quality. It also provided an increased specificity of 88.0 % compared to 81.7 %, showing that a greater proportion of the segments identified as low quality were indeed of low quality. Secondly, the ImP SQI provided more accurate and precise reference RRs than the agreement SQI. This is shown by a bias (95 % confidence interval, CI) of 0.0 (-0.1 - 0.1) bpm for the ImP SQI, compared to -1.1 (-1.7 - 0.6) bpm across all settings. This indicates that there was no systematic mean error between the reference RRs calculated by the ImP SQI and the manually determined reference RRs. In contrast, the agreement SQI tended to under-estimate the manually determined reference RRs by 1.1 bpm. In addition, the precision (2SD) of the ImP SQI was 1.9 (1.4 - 2.4) bpm, compared to 7.9 (5.8 - 9.9). Since 95 % of the errors are expected to lie within $\pm$ 2SD of the bias, a smaller precision indicates improved performance [309]. These results are in keeping with previous work, which found that time domain techniques should be used to obtain reference RRs from respiratory signals [109]. These results compare favourably with previous analyses of RRs obtained from the ImP signal, which have found that ImP-derived RRs can have 2SD values of $>$ 8 bpm [32].

The results suggest that the performance of the ImP SQI is not only an improvement on the previous methodology, but also sufficient for the assessment of RR algorithms. This is shown by a reasonably high proportion of high quality windows (74.0 %) being correctly identified. In addition, the reference RRs obtained from those windows identified as high quality were highly accurate and precise. The CP_2 and iCP_5 metrics demonstrate this further. The ImP SQI achieved a CP_2 of 97.8 %, indicating the proportion of reference RRs obtained which had an error of $<$ 2 bpm. In addition, it achieved an iCP_5 of 0.0 %, indicating that no reference

TABLE 4.14: The performances of the novel impedance pneumography (ImP) signal quality index (SQI), and the agreement SQI. Performances were assessed across the three clinical settings in the RRest Dataset. Discriminatory performance was assessed using the sensitivity and specificity of the SQI for identification of high quality signal segments. The performance of the SQI for estimation of RRs was assessed using the bias and limits of agreement (2SD), coverage probability (CP_2, proportion of values within 2 bpm of the reference value), and inverse coverage probability (iCP_2, proportion of values differing by more than 5 bpm from the reference value). Definition: CI - confidence interval

ImP SQI

Performance metric	Clinical setting			
	All	Ventilated in ICU	Unassisted in ICU	Shortly before ambulatory ward
	Discriminatory performance			
Sensitivity [%]	74.0	70.9	73.5	76.7
Specificity [%]	88.0	85.5	92.6	84.9
	RR estimation performance			
Bias [bpm] (95% CI)	0.0 (-0.1 - 0.1)	0.0 (-0.3 - 0.2)	0.0 (-0.2 - 0.1)	0.1 (-0.3 - 0.4)
2SD [bpm] (95% CI)	1.9 (1.4 - 2.4)	1.7 (0.9 - 2.5)	1.2 (0.6 - 1.7)	2.4 (1.3 - 3.6)
CP_2 [%]	97.8	97.9	100.0	95.9
iCP_5 [%]	0.0	0.0	0.0	0.0

Agreement SQI

	Discriminatory performance			
Sensitivity [%]	48.6	45.2	57.0	45.0
Specificity [%]	81.7	81.0	85.2	77.4
	RR estimation performance			
Bias [bpm] (95% CI)	-1.1 (-1.7 - -0.6)	-1.7 (-3.2 - -0.3)	-1.2 (-2.3 - -0.2)	-0.7 (-1.4 - 0.1)
2SD [bpm] (95% CI)	7.9 (5.8 - 9.9)	10.6 (5.5 - 15.7)	7.8 (4.1 - 11.4)	5.2 (2.7 - 7.6)
CP_2 [%]	79.5	71.4	75.2	89.1
iCP_5 [%]	3.8	7.1	3.8	1.6

RRs had an error of > 5 bpm. The robustness of these analyses are demonstrated by the total number of high quality windows over which the reference RRs were assessed, which was 407 for the ImP SQI. Given its high performance, the ImP SQI will allow more robust conclusions to be drawn from assessments of the performance of RR algorithms. It will allow differences between RRs estimated using ECG- or PPG-based RR algorithms, and reference RRs, to be largely attributed to poor RR algorithm performance, rather than poor reference RRs.

4.6.3 A Toolbox of RR algorithms

A key shortcoming of previous research into RR algorithms is the lack of benchmark implementations of algorithms. A toolbox of RR algorithms was designed and implemented to facilitate comprehensive assessments of RR algorithms. Similarly to the novel datasets described above, this toolbox has been used for the studies presented in this thesis, and is being made publicly available at http://peterhcharlton.github.io/RRest ensuring that future researchers can use standardised implementations of RR algorithms.

The toolbox is designed to evaluate the performance of a wide range of RR algorithms on different datasets. It consists of five main parts:

- **RR algorithms:** Estimation of RRs from ECG and PPG signals using all possible combinations of the implemented techniques listed in Tables 4.4, 4.5, 4.7, and 4.8;
- **Reference RR estimation:** Estimation of reference RRs from either a reference respiratory signal, or annotations of breath timings, or simultaneous RR measurements.
- **Signal quality assessment:** Assessment of the quality of ECG and PPG signals using the algorithms described in Section 3.3;
- **Respiratory signal quality assessment:** Assessment of the quality of respiratory signals extracted from ECG and PPG signals;
- **Statistical evaluation:** Evaluation of the performance of RR algorithms using a range of statistics.

The toolbox is designed to meet the wide range of needs of future researchers. It can be used with a range of datasets, as it is compatible with all of the novel datasets described in this chapter, and is also accompanied by scripts to re-format the MIMIC-II, CapnoBase, and

Fantasia datasets for compatibility. This facilitates multi-dataset assessments of RR algorithms covering a range of patient populations. Most parts of the toolbox contain implementations of several analytical methods, allowing the user to use methods of their preference. For instance, a total of 31 interchangeable techniques have been implemented for use at different RR algorithm stages. Reference RRs can be estimated using a range of methods, including the agreement SQI and novel ImP SQI described above. Three methods for assessment of the quality of respiratory signals are provided. Five statistical metrics are automatically calculated for users to evaluate algorithms. Researchers wishing to use the toolbox are directed towards a tutorial written to allow the novice to quickly start using it [6], and the manual available on the project website. Contributions of additional algorithms are welcomed.

The implementations of the 370 toolbox algorithms were assessed using the RRest-synth Dataset. Each technique implemented for extraction of respiratory signals (X_{A1} to X_{A4} and X_{B1} to X_{B10}), and for estimation of RRs (E_{F1} to E_{F7} and E_{T1} to E_{T5}), was considered to be successfully implemented if over half of the algorithms containing that technique were *acceptably accurate*. *Acceptably accurate* was defined as an absolute error of ≤ 1 bpm for at least 50 % of the simulated signals for any of the three modulations, on either the ECG or PPG. The threshold of 50 % was chosen to ensure that algorithms which performed well for a subset of the trialled HR and RR combinations were not excluded.

The E_{F5} and X_{B10} techniques were found not to be acceptably accurate. E_{F5} performed poorly at RRs outside of 12 - 20 bpm, which may be due to its bias towards identifying lower frequencies as the RR. X_{B10} variably detected the end of the PPG pulse as either the time of the minimum immediately before the diastolic peak, or the time of the diastolic peak, causing inaccuracies. After exclusion of these techniques 314 algorithms remained for further analyses.

4.7 Final Remarks

Many wearable sensors do not currently monitor RR, despite its importance for early detection of deteriorations. One potential solution is to estimate RR from the ECG or PPG, both of which are commonly monitored by wearable sensors. In this review a systematic search strategy was used to identify 140 publications describing algorithms to estimate RR from the ECG and PPG. The review provides a summary of the proposed algorithms and a synthesis of the methodologies used to assess them. The results led to the identification of three key limitations of previous

assessments of RR algorithms for assessing their potential utility for incorporation into wearable sensors to monitor acutely-ill patients:

- Previous assessments of RR algorithms have mostly been conducted on data acquired from populations which cannot be considered to be representative of acutely-ill patients;
- The reliability of reference RRs used in most assessments was not assessed, meaning the performances of RR algorithms were potentially reported as being poorer than they actually were; and
- No standardised implementations of algorithms are publicly available, making it difficult to conduct large-scale comparisons of algorithms.

These limitations were addressed by creating a toolbox of resources. The toolbox consists of: novel datasets including some from acutely-ill patients; a novel signal quality index for obtaining reference RRs from the ImP signal; and a toolbox of benchmark RR algorithms. These resources will be used to conduct analyses in the remainder of this thesis, and are being made publicly available for other researchers at: http://peterhcharlton.github.io/RRest.

Chapter 5

Extracting Respiratory Signals from the Electrocardiogram and Photoplethysmogram: Technical and Physiological Determinants

This chapter presents a study to assess how technical and physiological factors influence the quality of respiratory signals extracted from the electrocardiogram (ECG) and photoplethysmogram (PPG). Quality was assessed using the correlation between an extracted respiratory signal and the reference respiratory signal. The following technical factors were assessed: PPG measurement site (either the ear or finger); signal acquisition equipment type (either high fidelity laboratory equipment, or routine clinical equipment); type of input signal (ECG or PPG); and, signal sampling frequency. The following physiological factors were assessed: age, gender, respiratory rate (RR) and heart rate (HR). Recommendations based on the results are provided regarding wearable sensor designs for RR estimation, and clinical applications.

5.1 Introduction

A fundamental step in estimation of RR from the ECG and PPG is the extraction of a respiratory signal: a signal dominated by respiration. In the review of RR algorithms (Section 4.4.3), it was found that respiratory signals can be extracted from the ECG and PPG using either feature- or

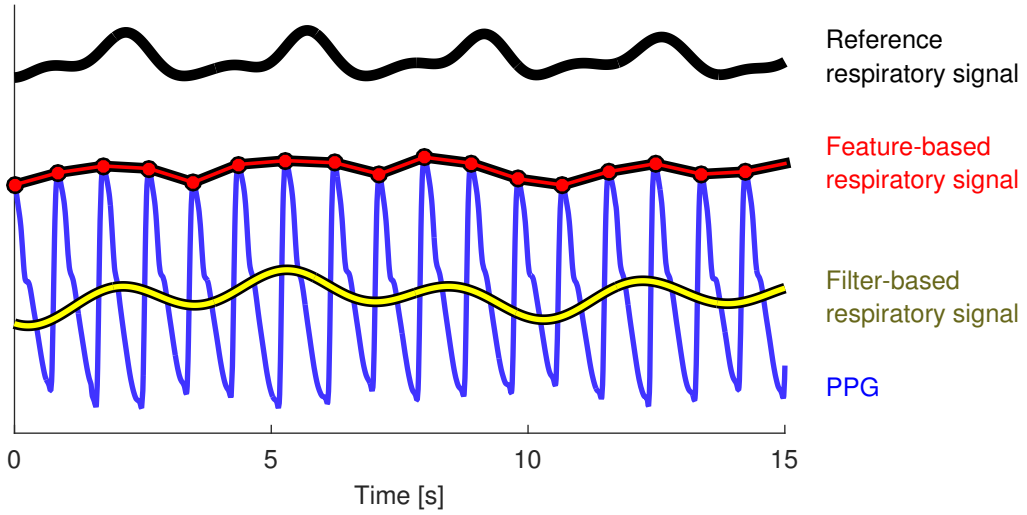

FIGURE 5.1: Extraction of respiratory signals using exemplary feature- and filter-based techniques. Respiratory signals have been extracted from the PPG using a feature-based technique in which pulse peak amplitudes are extracted, and a filter-based technique using the amplitude of the continuous wavelet transform. In this study the quality of extracted respiratory signals was assessed by calculating the correlation between each extracted respiratory signal and the reference respiratory signal. Source: [7] (CC BY 3.0, DOI: 10.1088/1361-6579/aa670e)

filter-based techniques, as illustrated in Figure 5.1. The processes for extraction of respiratory signals are demonstrated in Figure 5.2. This figure illustrates extraction of respiratory signals for each of the three idealised types of respiratory modulation of the ECG and PPG: baseline wander (BW), amplitude modulation (AM), and frequency modulation (FM) [10]. If the amplitude of the respiratory signal is too small compared to underlying noise, then the signal may not be distinguishable from the noise, preventing the precise estimation of RR. Thus, any factors which reduce the amplitude of respiratory modulations may result in reduced respiratory signal quality, affecting the estimation of RR from these signals.

The aim of the study presented in this chapter was to determine how the quality of respiratory signals is affected by technical and physiological factors which may be encountered in the clinical setting. Technical factors are those which are fixed during device design, such as the choice of anatomical site at which to measure the PPG. It is important to understand the influence of technical factors to optimise device design. In contrast, physiological factors cannot be controlled for. The influences of physiological factors, such as age, can inform decisions on whether or not particular RR algorithms are appropriate for use in wearable sensors for monitoring acutely-ill patients. Quality was measured using the correlation between an extracted respiratory signal and a reference respiratory signal (see Figure 5.1).

The chapter is structured as follows. Firstly, a review is presented of previous investigations

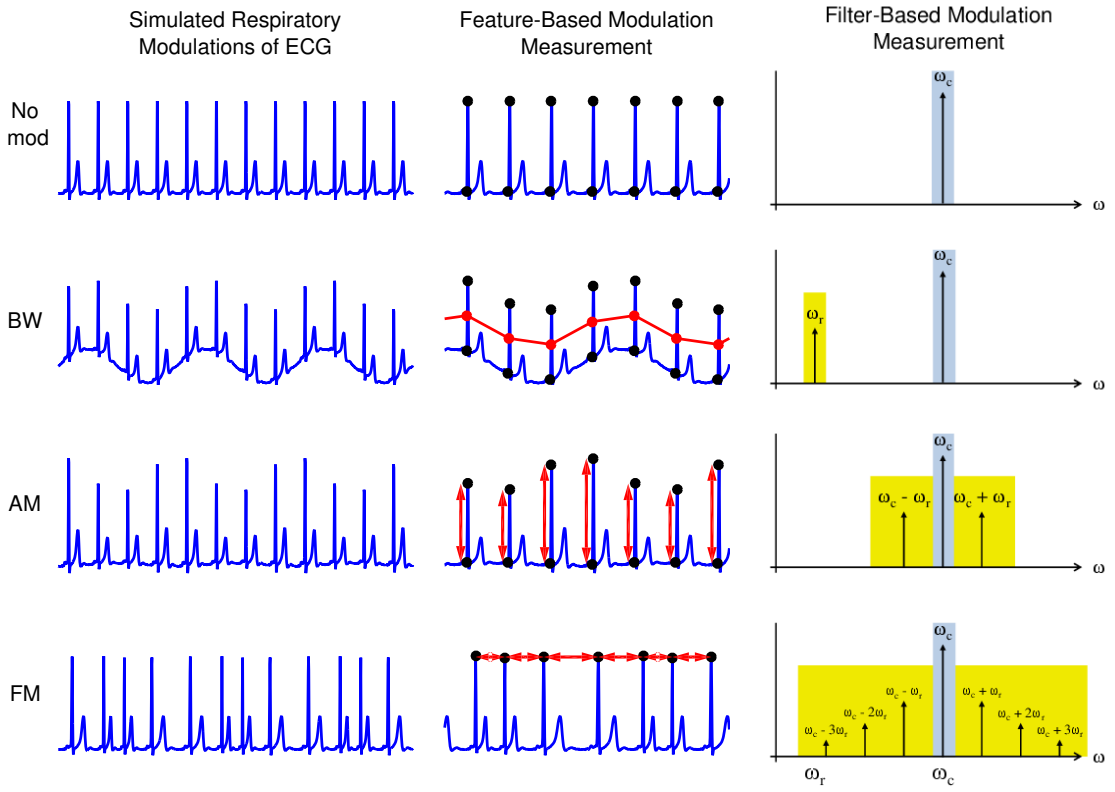

FIGURE 5.2: Processes for extraction of respiratory signals. On the left from top are shown simulated ECG signals with no modulation, baseline wander (BW), amplitude modulation (AM), and frequency modulation (FM). In the central column the Q- and R-waves have been identified (shown as dots), allowing feature-based modulation measurement of BW, AM and FM (shown in red). On the right are the corresponding frequency spectra of idealised signals at the cardiac frequency (ω_c) under the influence of each modulation. Filter-based modulation measurement consists of extracting signals dominated by the respiratory frequency. Note that only BW is manifested in the respiratory frequency (ω_r) band. Source: [7] (CC BY 3.0, DOI: 10.1088/1361-6579/aa670e)

into the influence of technical and physiological factors on respiratory signals of the ECG and PPG, and the subsequent performance of RR algorithms. In Section 5.3, the methods are described for data collection, assessment of the quality of respiratory signals, and statistical analysis. The results are then presented for each technical and physiological factor in turn. In Section 5.5, the impacts of these factors on device designs and on the use of RR algorithms in particular clinical scenarios are discussed. The RRest-healthy Dataset, respiratory signal extraction algorithms, and analysis code used in this study are being made publicly available at: `http://peterhcharlton.github.io/RRest`.

Postscript: A modified version of the study presented in this chapter has been published in [7]. The methodologies differ slightly because: (i) the analyses of physiological factors were also conducted using ECG signals from the Fantasia dataset in this version; and, (ii) a few

additional techniques were used to extract respiratory signals in [7]. Much of the content of this chapter has been adapted or reproduced accordingly from [7] (, DOI: 10.1088/1361-6579/aa670e).

5.2 Review of Previous Work

The previous work relating to each of the factors assessed in this study is now reviewed.

5.2.1 Technical factors

PPG probes can be positioned at a range of anatomical sites, including the finger, ear, forearm, shoulder and forehead [319]. Of these, only finger and ear measurements are widely used in clinical practice. The quality of respiratory signals extracted from the PPG may differ at different sites because of the augmentation of the systolic portion due to arterial pressure wave reflections [320], and the visco-elasticity of the arterial system [321]. Indeed, previous investigations have shown that the amplitude of BW is greater when the probe is positioned at the ear than the finger [319], [322]. However, further investigation is required to verify this finding and determine the effect of measurement site on AM and FM signals. This may impact device designers' considerations of the site of PPG measurement for RR estimation.

The equipment used to acquire ECG and PPG signals may influence the quality of respiratory signals. This is of particular concern with the PPG, since clinical monitors commonly output a filtered version which has been optimised for display, which may differ from the measured signal [323]. The processing procedures include auto-gain, auto-centre, and amplitude gain functions [324]. These adaptive filters may function over a short time scale, comparable to that of breathing, therefore potentially affecting extracted respiratory signals. Indeed, a recent study reported that the AM signals extracted from PPG signals acquired from two clinical monitors were not interchangeable [325]. Since monitors' filtering characteristics are not usually published [323], it is not clear how extracted respiratory signals are affected by this process. If high-fidelity laboratory equipment results in higher quality respiratory signals than a clinical monitor, then device designers may need to consider modifying the hardware in devices in order to extract high quality respiratory signals prior to filtering.

The type of input signal, ECG or PPG, may impact the quality of respiratory signals since different physiological mechanisms cause the respiratory modulations in the ECG and PPG. Therefore, the strengths of individual modulations may differ between the two signals, impacting extracted respiratory signals. The physiological mechanisms have been reported previously in [79] and [140], and were summarised in Table 4.1 (Section 4.2), although they are not fully understood. FM-based RR algorithms have previously been found to perform better when using the PPG rather than the ECG [326], [172], [191]. Further research is required to determine whether one signal is superior to the other for measurement of respiratory signals.

The sampling frequency of the input signal may affect the quality of respiratory signals. This is most important for the ECG signal since many of the feature-based respiratory signals are calculated from measurements of the QRS-spike, which contains high frequency content. It is intuitively appealing to use high sampling rates to ensure that respiratory modulations are captured as precisely as possible. Indeed, several studies have used high-fidelity equipment sampling the ECG and PPG at up to 1 kHz [327], [328]. However, it is desirable to be able to use low fidelity equipment since it will make ECG- and PPG-based RR estimation more widely accessible, particularly in resource-constrained settings. For instance, smart phones with PPG sampling rates as low as 30 Hz [329] are widely accessible. However, any reduction in signal quality due to lower sampling frequencies must be appreciated to allow appropriate equipment to be selected for each clinical setting.

5.2.2 Physiological factors

Age may affect the quality of respiratory signals since some of the physiological mechanisms which cause respiratory modulations of the ECG and PPG diminish with age. In particular, respiratory sinus arrhythmia (RSA, which causes FM) and chest wall expansion (which is linked to BW and AM) both diminish with age [141], [142], [143], [10]. Indeed, FM-based ECG algorithms of RR have been found to perform worse in older subjects [168], [109], [256], [237]. However, a previous investigation into the effect of age on BW-based PPG algorithms of RR did not find a difference in performance with age [235]. Further investigation is required to determine the extent to which each respiratory modulation is affected by age. This is particularly important given that many acutely-ill hospital patients are elderly, as shown by a median age of patients in the LISTEN Dataset of 67.5 years (Chapter 3).

It has also been suggested that gender may influence the quality of respiratory signals. The amplitude of FM in the PPG has been observed to be greater in women than men [330]. In contrast, the amplitude of BW in the PPG does not appear to be influenced by gender [235]. If the qualities of respiratory signals differ between women and men then potentially different respiratory signals could be extracted for each gender.

It has also been reported that the amplitudes of respiratory modulations are affected by a subject's RR. This would be particularly significant if it results in a reduction in the performance of RR algorithms at abnormally low or high RRs, since it is important to be able to detect these extreme values to ensure patient safety [331]. RSA, the mechanism which causes FM, is reduced above a certain corner respiratory frequency [332]. Furthermore, it has been observed that AM of the PPG is reduced at increasing RRs [333]. It has been suggested that the reduced amplitude of respiratory modulations at elevated RRs causes a reduction in the performance of RR algorithms [188], [189], [334], [329], [327]. Another study found that FM-based ECG algorithms performed worse at higher RRs, whereas AM-based algorithms performed better at higher RRs [111]. It has also been proposed that there is a range of RRs within which RR algorithms perform best, and that performance is reduced for RRs outside of this range. However, the exact range is unclear, having being reported as 8-11 breaths per minute (bpm) [189], and 16-20 bpm [237].

5.3 Methods

The methods used for both data collection and signal processing have, in part, already been described in [10]. Those relevant to this study are presented here.

5.3.1 Technical and physiological factors

The technical and physiological factors investigated in this study are listed in Table 5.1. The investigations were carried out as follows. Firstly, the respective qualities of respiratory signals extracted from finger and ear PPG signals were compared. The measurement site associated with lower quality respiratory signals was eliminated from further analyses. Secondly, respiratory signals extracted from laboratory and clinical signal acquisition equipment were compared. Similarly, the signal acquisition equipment associated with lower quality respiratory signals was eliminated from further analyses. Finally, the influences of the remaining technical factors, and the physiological factors, on respiratory signal quality were assessed.

TABLE 5.1: Technical and physiological factors investigated in this study which may influence the quality of respiratory signals extracted from the electrocardiogram (ECG) and photoplethysmogram (PPG).

Technical	**Physiological**
PPG measurement site: finger or ear	Age
Signal acquisition equipment: laboratory or clinical	Gender
Input signal: ECG or PPG	Respiratory rate (RR)
Sampling frequency	Heart rate (HR)

5.3.2 Datasets

The primary dataset used in this investigation was the RRest-healthy Dataset [7]. This dataset is described in detail in Section 4.6.1. Briefly, it contains data acquired from 41 young healthy subjects (aged between 18 and 39 years old), and 16 elderly subjects (at least 70 years old), for approximately 10 mins at rest. ECG, PPG and reference oral-nasal pressure respiratory signals were acquired simultaneously. This dataset is suitable for the assessment of technical factors since it contains signals acquired using both laboratory and clinical equipment simultaneously, and using different PPG measurement sites. This dataset is also suitable for the assessment of physiological factors since it contains data from young and elderly, and male and female subjects.

The RRest-healthy Dataset was extended in two ways to allow the impact of sampling frequency to be assessed. Firstly, ECG and PPG signals were downsampled incrementally from the original sampling frequencies to 50 Hz for the ECG and 8 Hz for the PPG. The downsampled signals were then interpolated at the original sampling frequency using cubic-spline interpolation.

A secondary dataset, the publicly available Fantasia dataset [302, 303], was used to assess the generalisability of the results pertaining to a subset of technical and physiological factors. This dataset contains data acquired from 20 young healthy subjects (21 - 34 years old), and 20 elderly subjects (68 - 85 years old). The recordings contain simultaneous ECG and reference respiratory signals at 250 Hz. The first 66.7 mins of each subject's recording were used for analysis. This dataset is suitable for assessing the physiological factors considered in this investigation. However, it does not contain the range of signals provided by the RRest-healthy Dataset, so was not used to investigate technical factors.

Table 5.2: Demographic characteristics of the two datasets used in this investigation.

Characteristic	Young Cohort	Elderly Cohort
RRest-healthy Dataset		
No. subjects	41	16
Age [years], med (lower - upper quartiles)	29 (26 - 31)	75 (72 - 78)
Female [%]	51	56
Fantasia Dataset		
No. subjects	20	20
Age [years], med (lower - upper quartiles)	26 (22 - 30)	73 (71 - 77)
Female [%]	50	50

Table 5.3: Data characteristics

Characteristic	No. high quality 32 s windows per subject, med (lower - upper quartiles)
RRest-healthy Dataset	
Reference respiratory signal (ref)	19 (18 - 20)
Laboratory ECG and ref	19 (18 - 20)
Laboratory finger PPG and ref	18 (16 - 20)
Laboratory ear PPG and ref	19 (17 - 20)
Clinical ECG and ref	19 (18 - 20)
Clinical finger PPG and ref	19 (17 - 20)
Fantasia Dataset	
Reference respiratory signal (ref)	72 (39 - 96)
ECG and ref	56 (33 - 91)

The demographic characteristics of the subjects in each dataset are provided in Table 5.2. Data from each subject consisted of a median (lower - upper quartiles) of 20 (19 - 20) 32 s windows in RRest-healthy, and 124 (124 - 124) 32 s windows in Fantasia. The number of high-quality windows for each signal are given in Table 5.3. The reference values of RR in the RRest-healthy Dataset were 16.5 (11.6 - 19.7) bpm, and 17.5 (14.6 - 19.1) bpm in the Fantasia Dataset. The reference values of HR in the RRest-healthy Dataset were 64.0 (58.1 - 69.1) bpm, and 58.8 (55.0 - 63.4) bpm in the Fantasia Dataset.

5.3.3 Quality assessment

ECG, PPG and oral-nasal pressure signals were segmented into adjacent windows of 32 s duration. The quality of each signal during each window was assessed using the methods described below. Any windows in which any of the required signals were of low quality were excluded from analyses.

The quality of ECG and PPG signals was assessed using the algorithm reported in [124], and described in Section 3.3. The quality of the oral-nasal pressure signal in the RRest-healthy Dataset was assessed by calculating its signal-to-noise ratio using a modified periodogram. Any windows with a low signal-to-noise ratio were deemed to be of low quality, with the threshold for exclusion set to eliminate windows in which breaths could not be identified visually. The quality of the reference respiratory signal in the Fantasia Dataset was assessed using the SQI designed and validated in Section 4.6.2.

5.3.4 Extraction of respiratory signals

Several techniques have been proposed for extraction of respiratory signals from the ECG and PPG, as described in Section 4.4.3.1. In this study four filter-based techniques (X_{A1} to X_{A4}, described in Table 4.4), and nine feature-based techniques (X_{B1} to X_{B9}, described in Table 4.5) were used to extract a wide range of respiratory signals. All techniques were used with the RRest-healthy Dataset. However, techniques X_{A2} to X_{A4}, and X_{B9} were not used with the Fantasia Dataset since they were too computationally expensive to be used with the longer recordings in this dataset.

Filter-based techniques, X_{A1} to X_{A4}, were implemented as described in Table 4.4, followed by elimination of frequency content outside of the range of plausible respiratory frequencies by band-pass filtering.

Feature-based extraction was conducted as follows. Very high frequencies were eliminated using low-pass filters with -3 dB cut-offs of 100 and 35 Hz for the ECG and PPG, respectively. An additional 50 Hz notch filter was used to eliminate mains interference in the ECG. Beat detection was performed on the ECG using a QRS detector based upon the algorithm of Pan, Hamilton and Tompkins [72], [73], and on the PPG using the Incremental-Merge Segmentation (IMS) algorithm [74]. R-waves and pulse peaks were detected as the maxima at or between detected

beats. QRS troughs were detected as the minima within the 0.10 s prior to R-waves [112], and pulse troughs as the minima between pulse peaks [186]. One of the beat-by-beat features, X_{B1} to X_{B9}, was obtained as described in Table 4.5. Features derived from ectopic beats were eliminated using the algorithm described in [335]. The beat-by-beat features were generated at a variable rate (the heart rate). The time-series of these features was re-sampled at 5 Hz using linear interpolation since subsequent processing required a constant sampling frequency [80]. Frequency content outside of the range of plausible respiratory frequencies was eliminated by band-pass filtering.

The range of plausible respiratory frequencies was determined as follows. The lower cut-off was fixed at 4 breaths per minute (bpm). The upper limit was set to 36 bpm to bisect the maximum RR and minimum heart rate (HR) in the RRest-healthy Dataset (33 bpm and 40 bpm respectively). This limit was also suitable for the Fantasia Dataset since the maximum RR and minimum HR were 27 bpm and 40 bpm respectively) This ensured that the extracted respiratory signals were not contaminated with cardiac frequency content.

5.3.5 Respiratory signal assessment

The quality of extracted respiratory signals was assessed as follows. Signals were segmented into the 32 s windows defined during quality assessment. For each window, the extracted respiratory signals and simultaneous reference respiratory signal were re-sampled at 5 Hz using linear interpolation, band-pass filtered between 4 and 60 bpm to remove non-respiratory frequencies, and temporally aligned to account for any phase difference between the two signals. The quality of each extracted respiratory signal was calculated as the correlation between that extracted respiratory signal and the reference oral-nasal pressure signal (see Figure 5.1). The correlation was calculated using Pearson's linear correlation coefficient (CC) [330]. The remainder of the methodology varied according to the particular factor being investigated as follows:

- Comparisons between subjects (*e.g.* young vs. elderly subjects) were performed using subject-specific CCs. The subject-specific CC was found for each subject and for each respiratory signal by calculating the median of the extracted respiratory signal's CCs from each of a particular subject's windows.
- Comparisons between input signals (*e.g.* ear vs. finger PPG) were performed using subject-specific differences in CCs. The subject-specific difference in CCs was found for

each subject and for each respiratory signal by calculating the median difference between the CCs for the extracted respiratory signal when extracted from a first input signal, and the CCs for the extracted respiratory signal when extracted from a second input signal.

5.3.6 Statistical analysis

Statistical tests were performed using a significance level of $\alpha = 0.05$. The Wilcoxon signed rank test was used to compare simultaneously recorded signals, such as ear and finger PPGs. The Wilcoxon rank sum test was used for results from independent groups, such as those acquired from young and elderly subjects. When testing for trends, such as across a range of reference RRs, the Mann-Kendall monotonic trend test was used, as described in [336]. Kendall's rank CC was reported for statistically significant trends as an indicator of the strength of the trend, as described in [337]. This statistic indicates both the magnitude and direction of trends. A value of zero indicates no trend. Positive values indicates positive trends, and vice-versa, with values further from zero indicating stronger trends. The directionality of statistically significant differences was determined by using a normal approximation to compute a z-statistic corresponding to an approximate p-value, the polarity of which indicated directionality.

During the analysis of each factor, a statistical test was performed to identify any changes in the quality of each respiratory signal. Since 12 signals were tested for the ECG, and 10 for the PPG, this would usually increase the probability of a type I error (false rejection of a null hypothesis) considerably. Therefore, a Holm-Sidak correction was made to ensure that the probability of a type I error was fixed at 5% [338], [339].

Respiratory signals extracted from the ECG and PPG were ranked by identifying the signal with the greatest median CC (control), and assessing the probability that each other signal's CCs originated from the same distribution as the control signal.

5.4 Results

5.4.1 PPG measurement site

A total of six out of eleven respiratory signals had significantly greater CCs, indicating higher quality, when extracted from finger PPG signals than ear PPG signals. The remaining signals

showed no significant differences. The results of the comparisons between finger and ear PPG signals are provided in full in Table A.1 in the appendix.

5.4.2 Signal acquisition equipment

The respiratory signals extracted from laboratory and clinical signals were mostly comparable, with the quality of a minority of signals differing significantly in favour of one set of recording equipment. A total of three (out of 13) signals extracted from the ECG were of higher quality when using clinical equipment. When using the PPG, five signals were of higher quality when using clinical equipment, and two when using laboratory equipment (out of 11). Since neither set of recording equipment provided consistently higher CCs, only clinical signals were considered in the remaining comparisons to increase the clinical applicability of the conclusions. The results of the comparisons between signals acquired from laboratory and clinical equipment are provided in Table A.2 in the appendix.

5.4.3 Input signal: ECG or PPG

The subject-specific CCs of each respiratory signal extracted from the ECG and PPG are shown in Figure 5.3. All respiratory signals were ranked more highly when extracted from the ECG than the PPG. Indeed, all of the PPG-extracted respiratory signals had significantly lower CCs than ECG(X_{B2}), the ECG-extracted respiratory signal with the greatest median CC. Despite this, ECG and PPG signals were retained in the remainder of the analysis. This ensured the results were applicable to situations where device design considerations or clinical conditions enforce the use of one particular signal for practical, rather than performance-based, reasons.

These results also show that some techniques for extraction of respiratory signals performed particularly poorly, namely: X_{A4}, X_{B7} and X_{B9}. Furthermore, they indicate that feature-based techniques mostly performed better than filter-based techniques for extraction of respiratory signals (*e.g.* X_{B2} was ranked higher than X_{A2}). The results of specific comparisons between feature- and filter-based techniques for extraction of the three fundamental respiratory modulations are shown in Table A.3 in the appendix. Five out of six respiratory signals had significantly greater CCs when extracted using feature-based techniques than filter-based techniques.

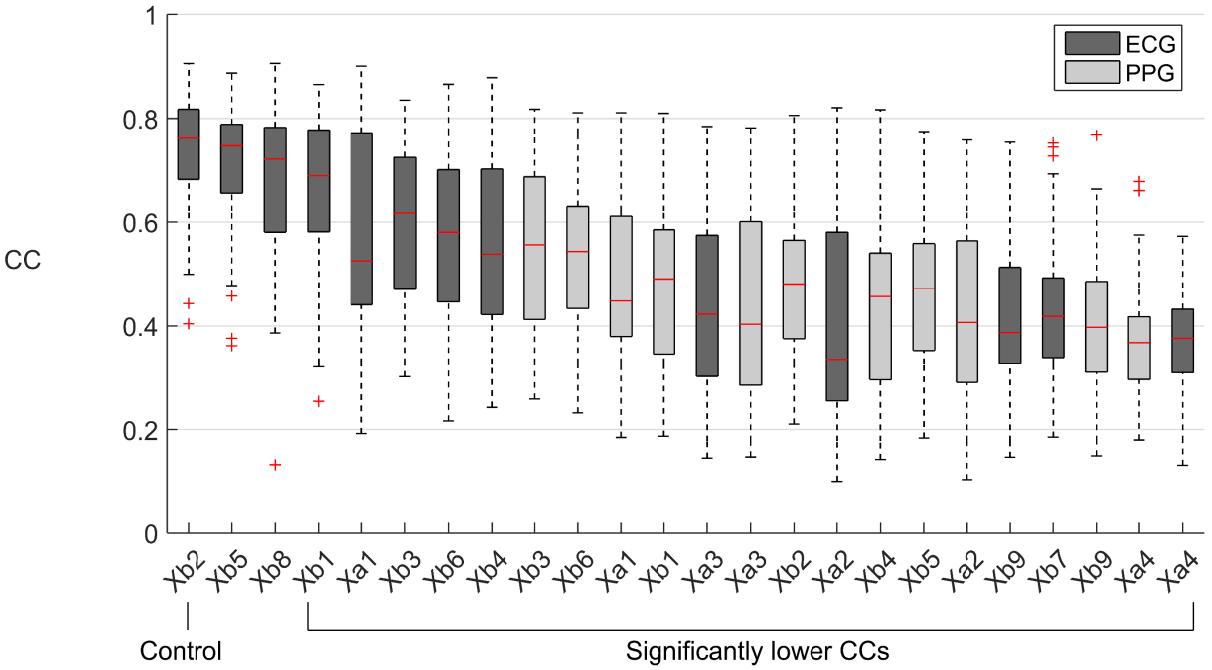

FIGURE 5.3: Comparison of the qualities of respiratory signals extracted from the ECG and PPG: Correlation coefficients (CCs) for each respiratory signal extracted from the ECG and PPG. The respiratory signal with the greatest median CC, ECG(X_{B2}), was used as a control against which each respiratory signal was compared in turn. Those with significantly lower CCs than this control are indicated. Note that all of the signals extracted from the PPG had significantly lower CCs than the best ECG-extracted signal. Outliers are shown by +.

5.4.4 Sampling frequency

Figure 5.4 shows the CCs of respiratory signals extracted from ECG and PPG signals at different sampling frequencies. Filter-based respiratory signals (X_{A1} to X_{A4}) were largely unaffected by sampling frequency. The CCs of all feature-based respiratory signals (X_{B1} to X_{B9}) extracted from the ECG except one were significantly lower at reduced sampling frequencies, beginning below 250 Hz. In contrast, CCs of feature-based respiratory signals extracted from the PPG were not reduced until the sampling frequency was below 16 Hz.

5.4.5 Age

In the results from the RRest-healthy Dataset, the CCs of PPG(X_{B3}), a respiratory signal based on FM, were significantly lower in elderly subjects than young subjects in the RRest-healthy Dataset. The CCs of other respiratory signals based on FM, namely PPG(X_{A3}), ECG(X_{A3}) and ECG(X_{B3}) were also substantially lower in elderly subjects, although these differences ($p = 0.04$, $p = 0.09$ and $p = 0.01$ respectively) did not reach statistical significance, partly due

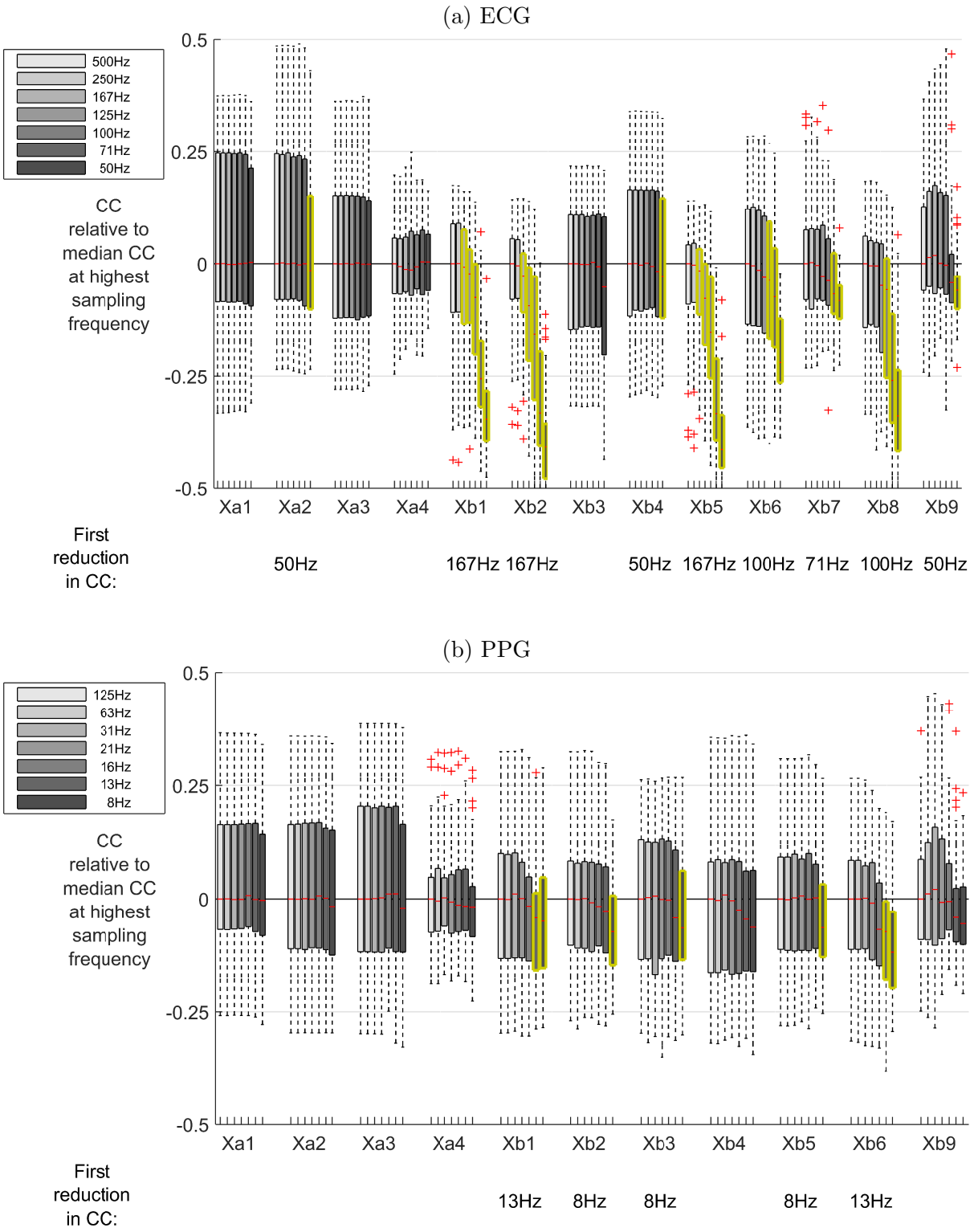

FIGURE 5.4: Comparison of sampling frequencies: Subject-specific correlation coefficients (CCs) are shown for each respiratory signal extracted from the ECG and PPG at different sampling frequencies. Significant reductions in CCs at lower sampling frequencies are highlighted in yellow. Filter-based respiratory signals (X_{A1} to X_{A4}) were largely unaffected by sampling frequency, whereas the CCs of many feature-based respiratory signals (X_{B1} to X_{B9}) were significantly reduced at lower sampling frequencies. Significant reductions in CCs of ECG-extracted respiratory signals occurred at sampling frequencies below 250 Hz, whereas reductions in CCs of PPG-extracted respiratory signals occurred below 16 Hz.

to the correction for multiple comparisons. For comparison, when the equivalent analysis was performed on laboratory signals, similarly only the CCs of PPG(X_{B3}) were significantly different (lower in elderly subjects).

In the results from the Fantasia Dataset, none of the CCs differed significantly between young and elderly cohorts. However, the CCs of the respiratory signal based on FM, ECG(X_{B3}), were similarly substantially lower in elderly subjects, although this difference ($p = 0.02$) also did not reach statistical significance due to the correction for multiple comparisons.

The results of comparisons between young and elderly subjects performed on the RRest-healthy and Fantasia Datasets are shown in Tables A.4 and A.5 respectively, in the appendix.

5.4.6 Gender

Similar sub-group analysis of male and female subjects were performed on the two datasets. No significant differences in quality were found for any respiratory signals between male and female subjects in either dataset.

5.4.7 Respiratory rate

The results of comparisons of respiratory signals at different reference RRs are shown in Figures 5.5 and 5.6 for the two datasets. The CCs of most respiratory signals extracted from the ECG, and all extracted from the PPG decreased significantly with increasing RR.

5.4.8 Heart rate

Only small, inconsistent changes in the qualities of respiratory signals were observed at different HRs. Since these were much smaller than the corresponding changes observed at different RRs, these were considered to be negligible within the range of HRs encountered in these datasets.

5.5 Discussion

A plethora of algorithms have been proposed for estimation of RR from the ECG and PPG by extracting a respiratory signal. In this study two datasets were used to investigate the effects

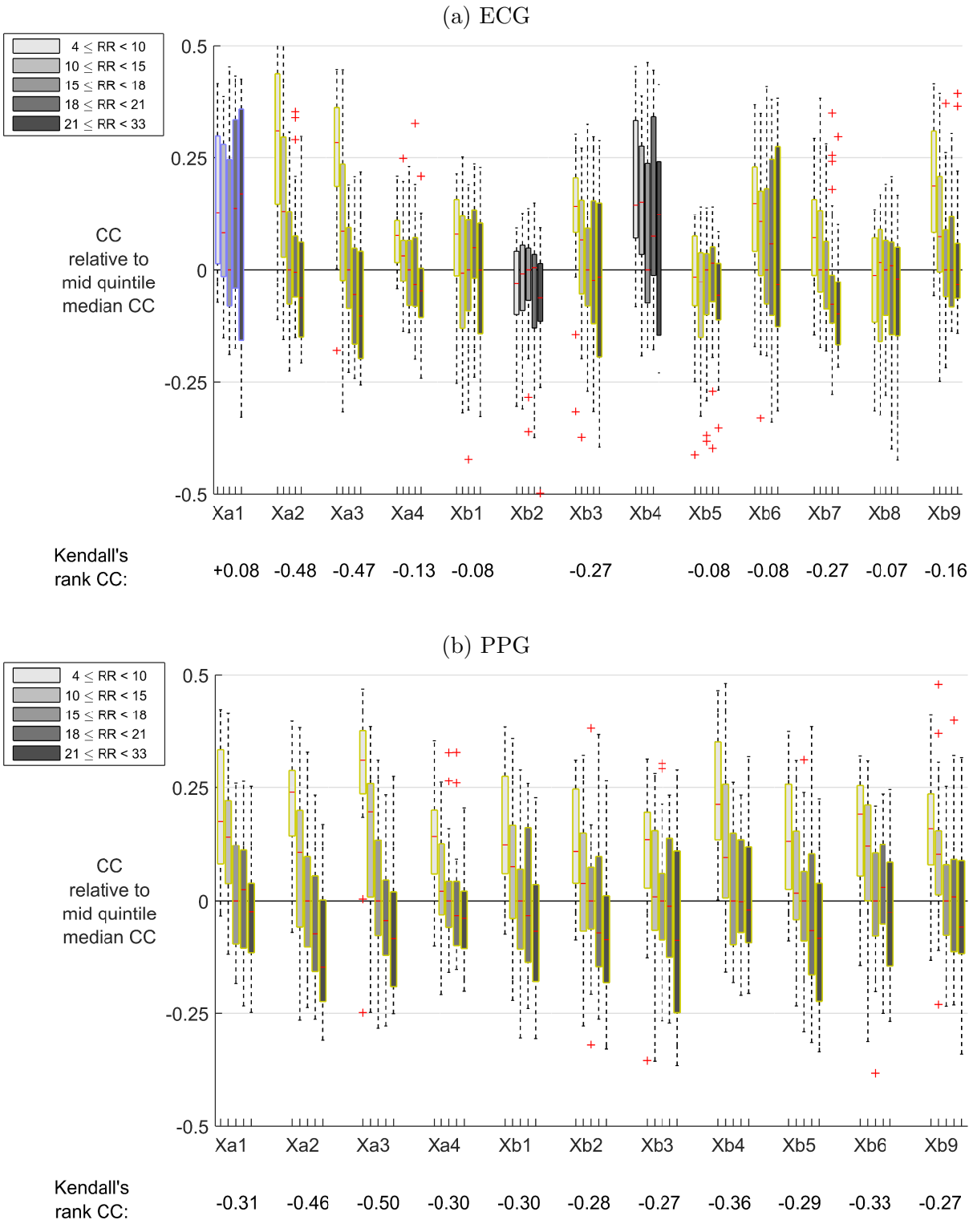

FIGURE 5.5: Trends in correlation coefficients (CCs) with respiratory rate (RR) in the RRest-healthy Dataset: The subject-specific CCs of respiratory signals obtained at different RRs are shown for (a) ECG, and (b) PPG. Significant trends indicating reduced CCs at increased RRs are highlighted in yellow, with blue indicating trends where CCs increased at increased RRs.

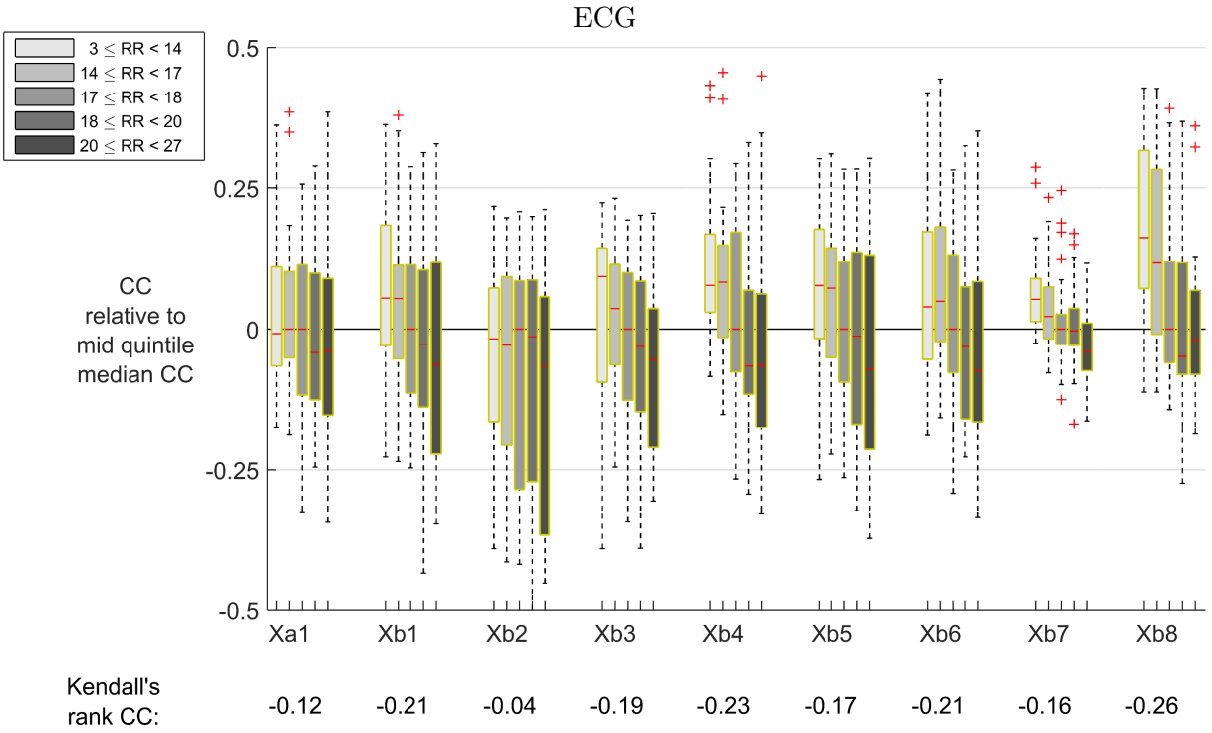

FIGURE 5.6: Trends in correlation coefficients (CCs) with respiratory rate (RR) in the Fantasia Dataset: The subject-specific CCs of respiratory signals obtained at different RRs are shown for the ECG. Significant trends indicating reduced CCs at increased RRs are highlighted in yellow.

of a range of technical and physiological factors on the quality of respiratory signals. The main conclusions are listed in Table 5.4, and are now considered in turn.

5.5.1 PPG measurement site

In current clinical practice the PPG is routinely measured at either the finger or ear for determining pulse rate and arterial blood oxygen saturation. Whilst cardiac modulation of the PPG remains strong across a range of anatomical sites, the respiratory modulation is affected by probe position [319]. The finding that many of the respiratory signals extracted from the finger PPG were of higher quality than those from the ear PPG, informs the recommendation that the finger PPG should be used in preference to the ear PPG for RR estimation. This was the case for many of the respiratory signals based on solely BW or FM: PPG(X_{A1}), PPG(X_{B1}), and PPG(X_{B3}). However, no significant differences were found in the qualities of respiratory signals based on AM: PPG(X_{A2}) and PPG(X_{B2}). This is in direct contrast to previous work, where BW was found to be stronger at the ear than the finger [319], [322]. Despite using the same transducers as in [322], the finger PPG tended to have a greater signal-to-noise ratio than the

TABLE 5.4: Main conclusions on the technical and physiological factors affecting the quality of respiratory signals extracted from the ECG and PPG.

Factor	Conclusion
	Technical factors
PPG measurement site	It is recommended that the PPG is measured at the finger rather than the ear for measuring respiratory signals.
Signal acquisition equipment	No evidence was found to suggest that high fidelity laboratory equipment provided superior measurement of ECG- and PPG-extracted respiratory signals when compared to a clinical monitor.
Input signal: ECG or PPG	The ECG should be used rather than the PPG for measuring respiratory signals where possible. Furthermore, the results suggested that feature-based techniques provide higher quality respiratory signals.
Sampling frequency	It is recommended that the ECG and PPG be measured at $\geq$ 250 Hz, and $\geq$ 16 Hz, respectively for measuring respiratory signals.
	Physiological factors
Age	It may be helpful to avoid the use of FM-based respiratory signals when estimating RR in elderly subjects.
Gender	There were no differences in the qualities of respiratory signals between women and men.
RR	Caution is recommended when using RR algorithms in settings where the detection of elevated RRs is required since the qualities of most respiratory signals were reduced at higher RRs.

ear PPG in our recordings (although it was not clear whether the cause of the lower signal-to-noise ratio at the ear was physiological or technical). This may explain why respiratory signals were of higher quality when extracted from the finger PPG in this study, whereas the ear PPG provided higher quality signals in previous studies.

5.5.2 Signal acquisition equipment

In this study the qualities of respiratory signals measured using laboratory and clinical equipment were compared to determine whether there were differences which prevented the use of

clinical monitoring signals for RR estimation. Counter-intuitively, more respiratory signals were of higher quality when measured using clinical equipment than when using laboratory equipment. This suggests that RR algorithms could be applied to this particular clinical monitor without a loss in performance, although this conclusion cannot be extrapolated to other clinical monitors which may contain different filtering procedures. This is in keeping with previously reported results [8].

5.5.3 Input signal: ECG or PPG

The vast majority of respiratory signals were ranked more highly when extracted from the ECG compared to the PPG. Therefore, it is recommended that where there is a choice of signals, the ECG should be preferred to the PPG. However, when taking individual measurements the PPG is often more convenient to acquire since a PPG finger probe can be more quickly attached than ECG electrodes. In contrast, previous results suggest that no trade-off is required between convenience and performance when performing continuous monitoring using wearable sensors, since in this setting ECG electrodes have been found to be better tolerated than a PPG probe [3].

A wide variation was observed in the qualities of respiratory signals extracted using different techniques. Consequently, six respiratory signal extraction techniques will be excluded from further analyses in this thesis. The following techniques were excluded due to poor performance: X_{A4}, X_{B7} and X_{B9}. In addition, the filter-based techniques X_{A1} to X_{A3} were excluded as they tended to be of lower quality than their feature-based equivalents.

5.5.4 Sampling frequency

It was hypothesised that the use of higher sampling frequencies would result in higher quality respiratory signals. In the ECG, this effect continued up to 250 Hz, above which there was no significant benefit to using higher sampling frequencies. This finding supports the use of RR algorithms with current clinical monitors since wearable sensors and static monitors typically sample the ECG at up to 256 Hz [75] and 500 Hz (this study). In the PPG, significant reductions in quality were only apparent below 16 Hz. This is promising since it suggests that non-specialist equipment, such as smart phones [200], tablets [329], and non-contact video cameras [139] should not be hindered by their relatively low sampling frequencies when measuring respiratory signals.

5.5.5 Age

Only one respiratory signal was of significantly lower quality in elderly subjects compared to young subjects in the RRest-healthy Dataset, whereas other signals showed no differences. This signal, PPG(X_{B3}), is FM-based. Other FM-based signals were consistently of lower quality in elderly subjects, although these reductions in quality were not statistically significant. This is in keeping with previous work, and informs the recommendation that FM-based respiratory signals should be avoided when using RR algorithms with elderly subjects. Previously, RR algorithms which fuse estimates from different respiratory signals have commonly included FM-based signals to increase precision [80] [237]. Further research is required to determine whether the performance of these fusion algorithms could be improved in elderly subjects by exchanging the FM-based input for an alternative respiratory signal.

5.5.6 Gender

The lack of differences between male and female subjects in this study indicates that gender is not an important factor in determining the quality of respiratory signals. This is supported by the relatively high number of subjects in each subgroup (30 female and 27 male in the RRest-healthy Dataset, and 20 female and 20 male in the Fantasia Dataset), suggesting that the lack of differences was not simply due to a lack of statistical power. Indeed, the present sample sizes are greater than in a previous analysis of differences due to gender (14 female and 14 male) [330].

5.5.7 Respiratory rate

In this study most respiratory signals were of lower quality at higher RRs, in keeping with previous work. This suggests that the performance of RR algorithms may be gradually reduced as the true RR increases. This would be clinically significant since an elevated RR is a key marker of clinical deterioration [331]. If this translates into an unacceptable performance of RR algorithms at elevated RRs then this would severely limit their clinical utility.

A potential concern with this analysis of the effect of RR on respiratory signal quality is that the observed reduction in quality is due to inter-subject differences in quality. For instance, a latent subgroup of subjects with lower respiratory signal quality may also breathe at higher

RRs by coincidence, rather than there being a causal link between elevated RR and reduced quality.

5.5.8 Limitations

The key limitations to this study are as follows. Firstly, this study was conducted in a laboratory setting with healthy subjects. This allowed investigation of the influences of technical factors, age and gender, without common confounders such as the reduction in autonomic nervous system functionality associated with diabetes [340]. Further investigation is required prior to applying the findings to clinical settings. Secondly, a particular set of laboratory equipment, and a particular clinical monitor were used. Therefore, the findings regarding signal acquisition equipment may not be universally applicable to all manufacturers' equipment. Thirdly, a known statistical property of correcting for multiple comparisons is that the probability of a type II error (a false negative) is increased. Consequently, some smaller differences in signal qualities may not have been identified. Finally, the precise nature of the reference respiratory signal in the Fantasia Dataset is not known. In this study the SQI developed for an impedance pneumography (ImP) signal was used to assess its quality.

5.6 Final Remarks

This study assessed the impact of technical and physiological factors on the extraction of respiratory signals from the ECG and PPG. This was achieved through analysis of ECG, PPG and reference respiratory signals from young and elderly healthy subjects. The main technical recommendations for extraction of high quality respiratory signals were: (i) to measure the PPG at the finger rather than the ear; (ii) where possible, to use the ECG rather than the PPG; and, (iii) to use sampling frequencies of $\geq$ 250 Hz for the ECG, and $\geq$ 16 Hz for the PPG. The main clinical recommendations were: (i) to consider avoiding the use of FM-based respiratory signals in elderly subjects; and, (ii) to expect the qualities of respiratory signals to be reduced at higher RRs. These recommendations will be helpful to equipment manufacturers to inform the design of wearable sensors, and to clinicians for determining whether RR algorithms should be used in their particular setting.

The study demonstrated that the quality of respiratory signals is also dependent on the technique used for respiratory signal extraction. Three techniques, X_{A4}, X_{B7} and X_{B9}, performed

particularly poorly, so will not be considered further in this thesis. In addition, the remaining filter-based techniques, X_{A1} to X_{A3}, will be excluded from further analyses, since their feature-based equivalents, X_{B1} to X_{B3}, tended to be of higher quality.

Further work is required to determine whether additional factors encountered in clinical practice, such as pathological cardiovascular and respiratory changes, affect the qualities of extracted respiratory signals.

The RRest-healthy Dataset, code to download and format the Fantasia Dataset appropriately, respiratory signal extraction algorithms, and analysis code used in this study are being made publicly available at http://peterhcharlton.github.io/RRest. These resources allow future researchers to reproduce the analyses presented here, and to use the dataset for additional studies.

Chapter 6

Respiratory Rate Algorithms for the Electrocardiogram and Photoplethysmogram: Performance in Healthy and Hospitalised Subjects

This chapter presents a study of the performances of algorithms to estimate respiratory rate (RR) from the electrocardiogram (ECG) and photoplethysmogram (PPG) in healthy and hospitalised subjects. The performances of 95 RR algorithms were assessed using four datasets: two acquired from healthy subjects, and two from hospitalised patients. The aims of this study were to determine: how well RR algorithms perform in each setting; which algorithms provide the best performance; and, which input signal (ECG or PPG) provides the best performance.

6.1 Introduction

Over 100 algorithms have been proposed to estimate RR from the ECG or PPG [10]. However, the systematic review presented in Chapter 4 demonstrated that their performances have not yet been compared systematically. Previous comparisons of algorithms have been limited by comparing only a small number of algorithms. Approximately half (53.6 %) of the publications which have assessed the performance of an RR algorithm compared multiple algorithms (Section 4.4.4). However, the most algorithms compared in a single publication (outside of the

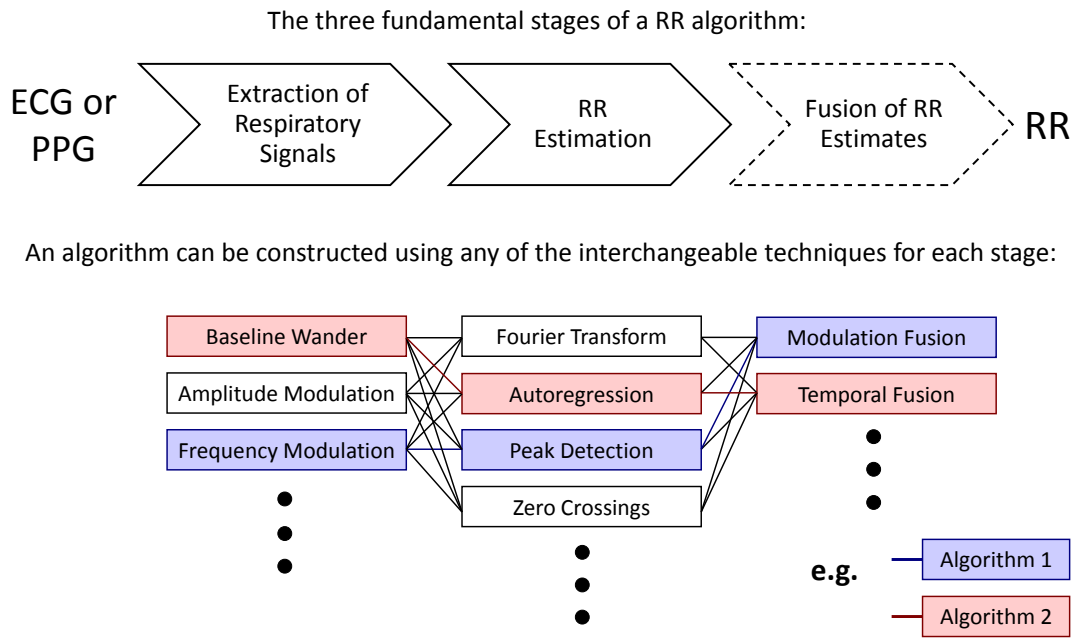

FIGURE 6.1: Construction of RR algorithms using different combinations of algorithmic techniques: RR algorithms can be constructed using any of the interchangeable techniques for each of the three fundamental stages. Two exemplary algorithms are shown. Adapted from [9] (CC BY 4.0, DOI: 10.5281/zenodo.439966).

current work) is 30, assessed using a single dataset [278]. This is not sufficient to determine which algorithm is the best, given that at least 49 different algorithmic techniques have been proposed for use in RR algorithms (Tables 4.4 to 4.8). Furthermore, several hundred algorithms can be constructed by trialling all possible combinations of the interchangeable techniques at each stage of RR algorithms, as demonstrated in Figure 6.1. Each algorithm may potentially have different performance characteristics. A convenient way to perform such a comparison could be to perform a meta-analysis of the existing literature. However, this is not practical for several reasons: (i) a wide range of incompatible statistics have previously been used to assess the performance of algorithms (Table 4.12); (ii) previous assessments of algorithms have used a wide range of datasets acquired from different patient populations, impeding comparisons; (iii) the complexity of RR algorithms means that algorithm implementations may differ considerably between publications, affecting the reported performances. Consequently, a novel study is required to perform a comprehensive assessment of RR algorithms, and to determine which (if any) algorithms perform sufficiently well for monitoring acutely-ill patients using wearable sensors.

Previous assessments of RR algorithms have suggested that improved performance can be achieved by using longer durations of input signals, and by fusing respiratory information from

different respiratory signals. Assessments of the effect of input signal duration have used durations of between 16 and 300 s [80, 108, 110, 341]. These assessments found trends towards improved performance when using longer durations, although there is no consensus on the optimal duration. Fusion of respiratory information from different respiratory signals has also been shown to improve performance [4, 80, 113, 237, 238, 341]. Usually information from two or three of the three fundamental respiratory signals (exhibiting BW, AM or FM) is fused, with a quality assessment procedure used to reject unreliable respiratory signals or RR estimates. However, there is similarly no consensus on the respiratory signals to be fused, nor the quality assessment procedure.

The aim of the study presented in this chapter is to perform a comprehensive assessment of the performances of RR algorithms in healthy and hospitalised subjects. Preliminary investigations were conducted to: (i) eliminate any algorithmic techniques which performed particularly poorly; (ii) to select an appropriate duration of input signal; and, (iii) to select the optimal configuration of fusion techniques. Following this, the performances of 95 RR algorithms were assessed on four datasets, consisting of two acquired from healthy subjects, and two from hospitalised patients. This assessment was used to: (i) quantify the algorithms' performances in these settings; (ii) identify the best-performing algorithms; and (iii) investigate how the input signal (ECG or PPG) affects performance. The toolbox of RR algorithms and the RRest-healthy dataset presented in Section 4.6 were used in this study. Both these resources and the analysis code are being made publicly available at: `http://peterhcharlton.github.io/RRest`.

6.2 Methods

6.2.1 Datasets

The four publicly available datasets summarised in Table 6.1 were used for this assessment. The first two datasets, the RRest-healthy [7] (described in Section 4.6.1) and Fantasia datasets [302, 303], contain data from both young and elderly healthy subjects (demographic characteristics provided in Table 5.2, Section 5.3.2). It is expected that RR algorithms will perform well on these datasets since they were acquired in laboratory conditions. The signals used from the RRest-healthy Dataset were the ECG (500 Hz) and PPG (125 Hz) signals acquired using a clinical monitor, and reference oral-nasal pressure respiratory signals (500 Hz). The signals used from the Fantasia dataset were the ECG and reference respiratory signals (both at 250

TABLE 6.1: The four datasets used for assessment of RR algorithms.

Dataset	Population	No. subjs	Age	ECG	PPG	Reference RRs
RRest-healthy	healthy	57	young and elderly adults	✓	✓	oral-nasal pressure
Fantasia	healthy	40	young and elderly adults	✓	✗	respiratory signal
MIMIC-II	critically-ill	127	adults	✓	✓	ImP signal
CapnoBase	elective surgery and anaesthesia	40	adults and paediatrics	✓	✓	manual annotations

Hz). Only the first 66.7 mins of each recording in the Fantasia dataset were used for this assessment, since data downloads from PhysioNet in Matlab ® format are limited to 1,000,000 samples. The remaining two datasets, MIMIC-II [299, 303] and CapnoBase [80], contain data from hospitalised patients. MIMIC-II contains data from adult patients in critical care. The signals used from the MIMIC-II dataset were the ECG, PPG, and impedance pneumography (ImP) respiratory signals (all at 125 Hz). A subset of MIMIC-II recordings from 127 ICU stays was used, each truncated at 30 mins to avoid excess data processing. CapnoBase contains data from both adult and paediatric patients undergoing elective surgery and routine anaesthesia [80]. Data from two subjects in the CapnoBase dataset were omitted as the recordings could not be analysed using the RRest Toolbox of algorithms, leaving data from 40 out of the original 42 subjects for analysis. The ECG and PPG signals in the CapnoBase dataset were sampled at 300 Hz, and the manual breath annotations were used to calculate reference RRs.

6.2.2 Quality assessment

Each subject's signals were segmented into adjacent windows for analysis. The quality of each signal during each window was assessed using the following methods. The quality of ECG and PPG signals was assessed using the template-matching signal quality index (SQI) algorithm reported in [124], and described in Section 3.3. The quality of the reference oral-nasal pressure respiratory signal in the RRest-healthy Dataset was assessed by comparing its signal-to-noise ratio to a threshold value, as described in Section 5.3.3. The quality of the reference respiratory signals in the Fantasia and MIMIC-II datasets were assessed using the novel SQI designed and verified in Section 4.6.2. Quality assessment was not required for a reference respiratory signal in the CapnoBase dataset since it was not used in the analysis. Any windows in which any

TABLE 6.2: Data characteristics: the number of 32 s windows per subject before and after quality assessment, median (lower - upper quartiles)

Dataset	No. windows per subject	No. high quality windows per subject		
		resp	resp & ECG	resp & PPG
RRest-healthy	20 (20 - 20)	20 (18 - 20)	20 (18 - 20)	19 (17 - 20)
Fantasia	124 (124 - 124)	71 (41 - 96)	56 (33 - 90)	n/a
CapnoBase	14 (14 - 14)	14 (14 - 14)	14 (14 - 14)	14 (14 - 14)
MIMIC-II	54 (45 - 56)	11 (3 - 24)	6 (0 - 21)	7 (2 - 17)

of the required signals were of low quality were excluded from the analysis. The numbers of windows included in the analysis are given in Table 6.2.

6.2.3 Respiratory rate estimation from the ECG and PPG

Algorithms to estimate RR from the ECG and PPG were constructed by selecting an interchangeable technique for each algorithm stage, as demonstrated in Figure 6.1. The following techniques were used at each stage:

- Extraction of respiratory signals: X_{B1} to X_{B6}, and X_{B8} (defined in Table 4.5);
- Estimation of RRs: E_{F1} to E_{F4}, E_{F6}, E_{F7}, and E_{T1} to E_{T5} (Table 4.7);
- Fusion of RRs (optional): F_{M1} to F_{M4} (Table 4.8), and F_{M5} (introduced in Section 6.3.3).

These techniques were chosen after elimination of techniques which performed particularly poorly in previous chapters (see Sections 4.6.3 and 5.6 for further details). Algorithms were constructed using all possible combinations of these techniques. Each algorithm was used with each window of input signal, outputting either a value for RR, or no output if the algorithm could not estimate RR from that window. The methods used at each stage of the RR algorithms are now described.

Respiratory signals were extracted from the ECG and PPG using the approach for feature-based techniques as described in Section 5.3.4. Previously, the range of plausible respiratory frequencies was previously set to 4 - 36 bpm to ensure that cardiac frequency content was excluded. However, in this assessment a wider range of 4 - 60 bpm was used to ensure that the RR algorithms were able to detect all RRs which may be encountered in clinical practice.

RRs were estimated from respiratory signals using time and frequency domain techniques. When using time-domain breath detection techniques (E_{T1} to E_{T5}) the estimated RR was calculated as the mean breath duration. When using frequency-domain spectral analysis techniques (E_{F1} to E_{F4}, and E_{F7}) the RR was identified as the frequency of the spectral peak with greatest magnitude within the range of plausible respiratory frequencies.

Finally, the optional fusion of RRs derived from BW, AM and FM respiratory signals (X_{B1} to X_{B3}) was performed using four modulation fusion techniques (F_{M1} to F_{M4}). F_{M1} can be used after any RR estimation technique, whereas the other modulation fusion techniques are restricted to use with particular RR estimation techniques. F_{M2} can only be used with frequency-domain RR estimation techniques since it requires a power spectrum, and F_{M3} and F_{M4} can only be used with E_{F4} since they utilise autoregressive modelling.

6.2.4 Optimisation of RR algorithms

Preliminary investigations were conducted using the RRest-healthy dataset to optimise the RR algorithms prior to assessment on the remaining datasets. Firstly, the performances of the algorithms were inspected to identify any techniques which performed particularly poorly. These techniques were excluded from further analyses. Secondly, three input signal durations were compared: 25, 32 and 50 s. To do so the irregularly sampled signals of beat-by-beat features were re-sampled at 5, 4 and 5 Hz respectively. This avoided excessive zero-padding of the signal when estimating RR from the signals using the fast Fourier transform. An input signal duration was selected for use in further analyses, based on RR algorithm performances and the need to obtain RR estimates frequently enough to detect deteriorations. Thirdly, the best-performing modulation fusion technique was optimised by trialling the use of either all three respiratory signals (X_{B1} to X_{B3}, corresponding to BW, AM and FM), or any two of these three signals. In addition, the quality assessment threshold was optimised. This novel configuration (F_{M5}) was then used as an additional technique in the analyses.

6.2.5 Reference respiratory rate estimation

When using the CapnoBase dataset, reference RRs were estimated from the manual breath annotations. When using the other datasets, reference RRs were estimated from the reference respiratory signals. For the Fantasia and MIMIC-II datasets, reference RR estimation was

performed as part of the reference respiratory signal quality assessment. This reference RR calculator has previously been found to have a *bias* of 0.0 bpm, and $2SD$ (standard deviation) of 1.9 bpm (Section 4.6.2; statistics explained below). For the RRest-healthy Dataset, a separate bespoke breath-detection algorithm was used to estimate reference RRs from the oral-nasal pressure signal. This algorithm consisted of the following steps:

- Band-pass filter the signal to eliminate non-respiratory frequencies (with 3 dB cut-off frequencies of 4 and 60 bpm).
- Normalise the signal within each window to have a mean of 0 and SD of 1.
- Detect inhalations as positive-gradient crossings of a specified threshold, k.

A threshold of $k = 0.42$ was determined by minimising the difference between the value of $2SD$ derived from the comparison of the reference RRs to expert breath annotations on 10 subjects. The performance of the reference RR calculator was then assessed against a second experts breath annotations on data from a different set of 10 subjects. It had a *bias* of 0.0 bpm and $2SD$ of 1.3 bpm.

6.2.6 Statistical analysis

Two approaches were used to assess the performance of RR algorithms. Firstly, several statistics were calculated to provide a thorough description of each algorithm's performance. Secondly, a cost function was designed in order to rank the algorithms according to their suitability for clinical use in wearable sensors.

The performance of RR algorithms was reported primarily using the limits of agreement (LoAs) technique [309] since it is widely used in comparisons of algorithms for estimation of biomarkers [342]. It consists of the following statistics:

- *bias*: The accuracy of RR algorithms was assessed by calculating their systematic bias [309],

$$bias = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}[RRest_{ij} - RRref_{ij}]\right)}{\sum_{i=1}^{n} m_i} \quad , \tag{6.1}$$

calculated as the mean difference between the estimated RR, $RRest$, and the reference RR, $RRref$, across the $i = 1, \ldots, n$ subjects, each of which had m_i pairs of estimated and reference RRs.

- $2SD$: The precision of RR algorithms was assessed by calculating the expected range of 95% of errors around the systematic bias, $\pm 1.96s$, denoted $2SD$ in this work [309]. The method described in [343] was used to account for repeated measurements per subject in the calculations. $2SD$ was calculated from the standard deviation of the errors, $s = \sqrt{\text{total variance}}$, where the total variance is the sum of the within-subjects and between-subjects variance. These two variance components were estimated using one-way analysis of variance:

$$\text{within-subjects variance} = MS_{residual} \quad , \quad \text{and} \tag{6.2}$$

$$\text{between-subjects variance} = \frac{MS_{subject} - MS_{residual}}{\frac{(\sum m_i)^2 - \sum m_i^2}{(n-1)\sum m_i}} \quad , \tag{6.3}$$

  where $MS_{residual}$ is the mean square error, and $MS_{subject}$ is the difference between the mean squares for subjects, and the sums are for $i = 1, \ldots, n$.

- $LoAs$: The limits within which 95% of errors are expected to lie, were estimated as [309]

$$95\% \text{ LoAs} = bias \pm 1.96s \quad . \tag{6.4}$$

The following statistics were also calculated to assess clinical utility:

- CP_2: The proportion of high quality windows for which highly useful $RRest$ values are returned, defined as being within ± 2 bpm of $RRref$. This statistic is known as the coverage probability, CP [344].

- iCP_5: The proportion of high quality windows for which potentially misleading $RRest$ values are returned, defined as being outside of ± 5 bpm of $RRref$. This statistic is referred to as the inverse coverage probability, iCP.

- TDI_{95}: The 95^{th} percentile of errors, indicating the error below which 95 % of errors fall. This statistic is referred to as the total deviation index, TDI [344].

- *prop*: the proportion of windows for which an algorithm provides a valid $RRest$ value, as opposed to a null output.

A cost function, C, was defined as:

$$C = \frac{CP_2}{iCP_5} \tag{6.5}$$

which varies between 0 and ∞, with higher values indicating improved performance. It is the ratio of CP_2 to iCP_5, and as such is the ratio of the probability that an $RRest$ value is within ± 2 bpm of $RRref$, to the probability that it is outside of ± 5 bpm of $RRref$. For example, $C = 10$ indicates that an $RRest$ value is ten times more likely to be within $\pm$ 2 bpm than to be outside $\pm$ 5 bpm. This cost function was used to rank algorithms to identify an algorithm for use with wearable sensors.

The Wilcoxon signed rank test was used to test for significant differences between the performances of algorithms when assessing paired data, and the Wilcoxon rank sum test was used when assessing non-paired groups.

6.3 Results

6.3.1 Elimination of techniques due to poor performance

Inspection of the performances of algorithms on the RRest-healthy Dataset revealed that those algorithms containing the E_{F4} technique performed poorly compared to the other algorithms, as illustrated in Figure 6.2. The technique, which uses auto-regressive all-pole modelling to estimate RR, was therefore excluded from further analyses. Consequently, no RR estimation techniques remained which use auto-regressive all-pole modelling, since the one other technique which uses this approach, E_{F5}, had been excluded previously due to poor performance on simulated data (Section 4.6.3). This resulted in the exclusion of two modulation fusion techniques, F_{M3} and F_{M4}, since they can only be used following RR estimation using an auto-regressive all-pole modelling technique.

6.3.2 Duration of input signal

The performances of the remaining algorithms for the three trialled input signal durations are shown in Figure 6.3. The performances of algorithms improved significantly when using longer durations of input signals, with a duration of 32 s providing improved performance over 25 s, and a duration of 50 s providing improved performance over both 25 and 32 s. Although performance

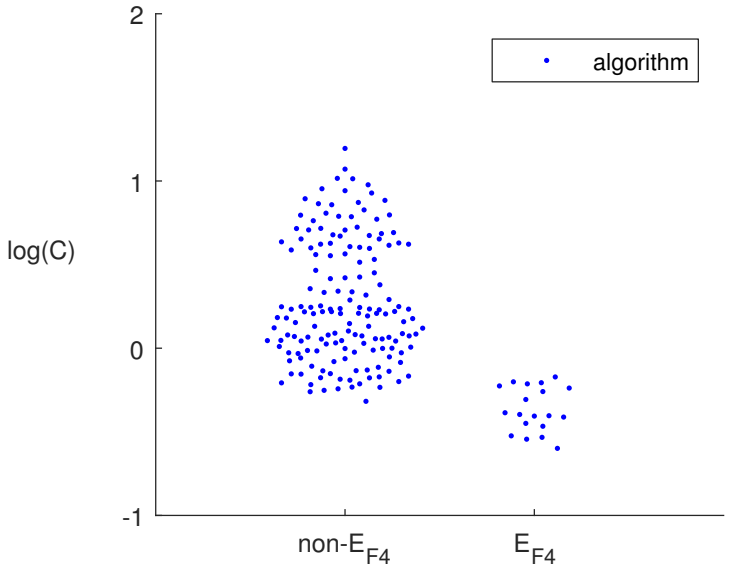

FIGURE 6.2: Poor performance of algorithms containing E_{F4}: This plot shows the performance of the algorithms on the RRest-healthy Dataset. The algorithms, each shown by a dot, have been categorised according to whether or not they used the E_{F4} technique. The E_{F4} technique was associated with poor performance as shown by the much lower C values for most of the algorithms containing it. These results were obtained using a 32 s duration of input signal, although similar results were obtained for other durations.

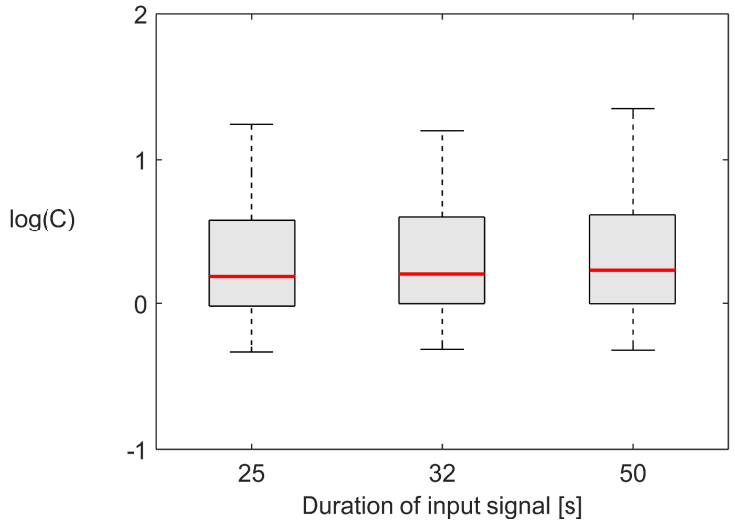

FIGURE 6.3: The performances of algorithms assessed on the RRest-healthy Dataset for input signal durations of 25, 32 and 50 s.

did improve when using longer durations, the size of the improvement was minimal, as shown by the median C values of: 1.55, 1.61 and 1.71 bpm for 25, 32 and 50 s respectively. Therefore, a duration of 32 s was chosen for the remainder of this work as a compromise between improved RR algorithm performance, and the need to maintain high data coverage rates (Figure 3.6, Section 3.4).

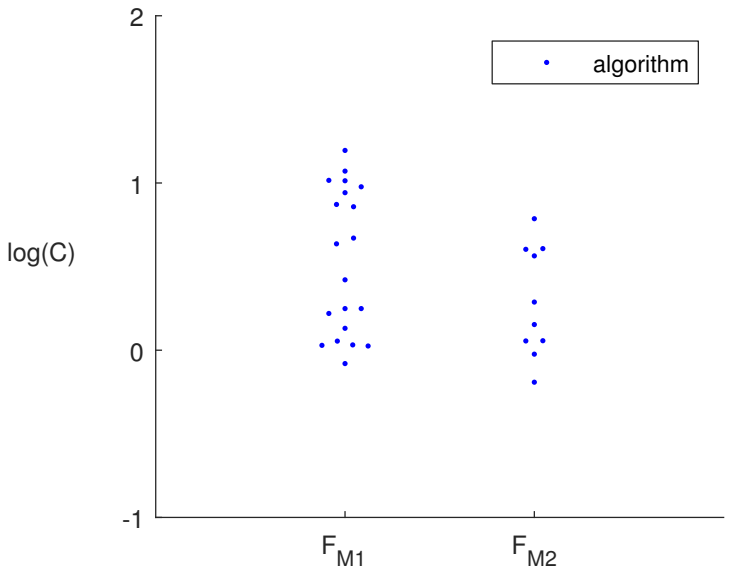

FIGURE 6.4: This plot shows the performances of algorithms which contained a modulation fusion technique on the RRest-healthy Dataset. The algorithms, each shown by a dot, have been categorised according to whether they used the F_{M1} or F_{M2} technique. The highest-performing algorithms used the F_{M1} technique.

6.3.3 Optimisation of fusion techniques

The performances of the algorithms which included a modulation fusion technique are shown in Figure 6.4. The highest-performing algorithms used F_{M1}. This suggested that this technique might provide the best overall performance across the multiple datasets under consideration. Therefore, an additional investigation was undertaken to optimise the performance of F_{M1} on the RRest-healthy Dataset.

Figure 6.5 shows how the performance of algorithms using F_{M1} varied as the input signals and value of the quality threshold were varied. The results corresponding to the ECG (first plot) suggest that performance could be improved by only using RR estimates derived from the BW and AM respiratory signals, since the BW, AM configuration appeared to give superior performance. In addition, the results demonstrate a compromise between the precision of RR estimates, and the proportion of windows for which RR estimates are provided. This compromise is determined by the quality threshold. A lower quality threshold tends to provide improved performance, albeit for a smaller proportion of windows. Based on these observations, a new configuration of the technique was proposed, consisting of using the BW, AM configuration and a quality threshold of 3.0. The performance of this new configuration is indicated by the hollow black dot. In contrast, the results corresponding to the PPG (second plot) do not provide rationale for modifying F_{M1}, since reducing the set of input signals did not provide a

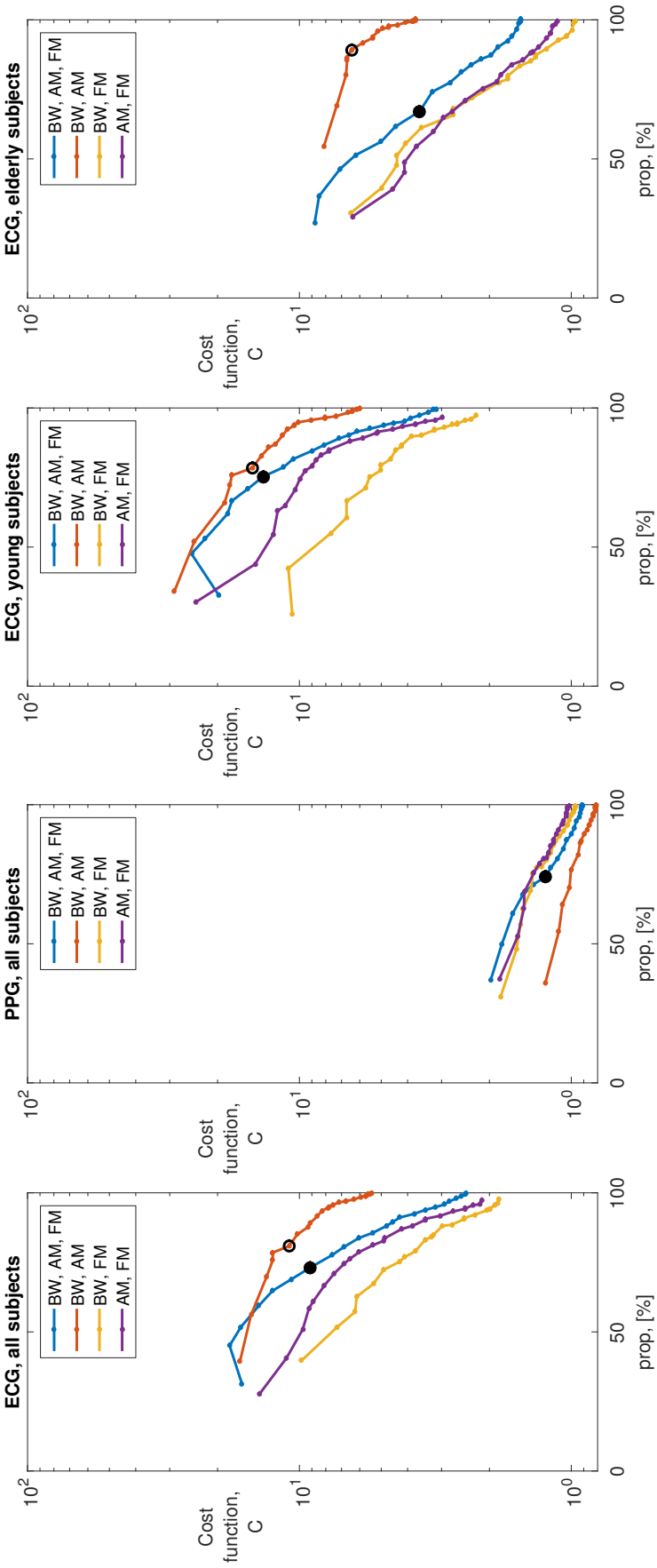

FIGURE 6.5: The performances of algorithms using variations of a modulation fusion technique, F_{M1}, are shown, assessed on the RRest-healthy Dataset. Performances are shown when algorithms were applied to (first plot) the ECG, (secondly) the PPG, (thirdly) ECG signals from young subjects, and (fourthly) ECG signals from elderly subjects. Performances are quantified using C, a bespoke cost-function indicating the clinical utility of an algorithm, and *prop*, the proportion of windows for which an RR estimate is returned (detailed in Section 6.2.6). Higher values of each indicate improved performance. The input signals to F_{M1} were varied, resulting in the four coloured lines on each plot. Each line shows the median algorithm performance when using RR estimates derived from either all three respiratory signals (BW, AM and FM), or two of the three. In addition the quality assessment threshold was varied from 0.5 to 12. Solid black dots indicate the performance when using the original F_{M1} implementation. Hollow black dots indicate the performance of the proposed adaptation for use with the ECG.

large increase in C, when compared to the original BW, AM, FM configuration. The third and fourth plots show a sub-group analysis of the results for the ECG for young (third plot) and elderly (fourth plot) subjects. These show that the new configuration of F_{M1} (indicated by a hollow black dot) provided higher performance than the original implementation (solid black dot) in both groups. However, the advantage was greatest in elderly subjects. Therefore, the new configuration of F_{M1} was used as an additional technique termed F_{M5} in the remaining analyses, with the expectation that it may provide superior performance when used with the ECG, particularly in elderly subjects.

6.3.4 Algorithm performances in health

The performances of 95 algorithms were assessed, ten of which contained X_{B8}. Those containing X_{B8} could only be used with the ECG signal since this technique calculates the area under the QRS complex.

The algorithms exhibited a wide range of performances in health, as demonstrated in Figure 6.6. The values of $2SD$ ranged from 3.6 bpm at best to 12.5 bpm at worst. Similarly, there were large variations in the other performance characteristics, with CP_2 ranging from 25.1 % to 82.7 %, iCP_5 from 2.2 % to 61.9 %, C from 0.4 to 36.6, and TDI_{95} from 3.7 bpm to 14.2 bpm.

Table 6.3 shows how the performance of algorithms varied according to whether a time- or frequency-domain RR estimation technique was used. Most performance statistics showed that algorithms which used time-domain RR estimation techniques performed significantly better than those which used frequency-domain techniques. Indeed, the use of time-domain techniques resulted in substantially higher C values than frequency-domain techniques when estimating RRs from ECG signals in the Fantasia dataset, and PPG signals in the RRest-healthy Dataset. Despite this apparent improvement in performance when using time-domain techniques, CP_2 and *prop* were not significantly improved by the use of time-domain techniques: CP_2 was either similar between the groups of algorithms, or improved in the frequency-domain group; and there were no significant differences in *prop* between the two groups. The reason for the difference in performance becomes apparent upon inspection of the differences in the distributions of RR algorithms errors, as shown in Figure 6.7. This figure shows that frequency-domain techniques were more vulnerable to underestimating RR. This is shown by the increased proportion of estimates resulting in errors of between -5 and -20 bpm. This tendency of frequency-domain techniques to underestimate RR explains why algorithms using time-domain estimates produced

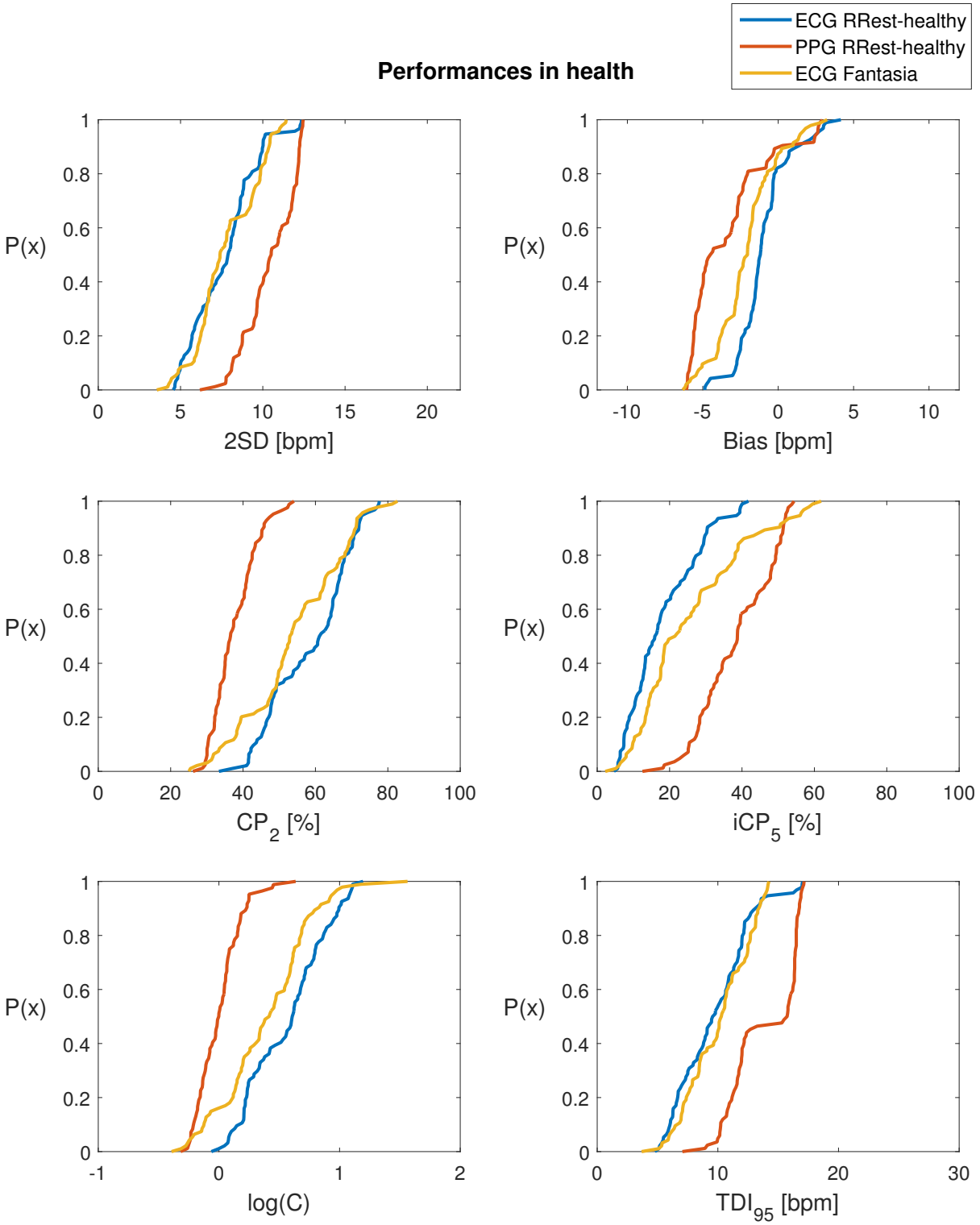

FIGURE 6.6: Performances of RR algorithms in health: Cumulative distribution functions of the performance characteristics of the algorithms assessed on the ECG and PPG signals in the RRest-healthy and Fantasia datasets (see legend).

TABLE 6.3: Performances of algorithms using time- and frequency-domain RR estimation techniques in health. Significant differences are annotated by *, with the domain providing improved performance labelled.

		Algorithm performance, med (quartiles)		
Signal	Statistic	Time domain	Freq domain	
ECG RRest-healthy	C	4.3 (1.9 - 7.4)	4.0 (1.7 - 5.9)	
	$2SD$	6.8 (5.7 - 8.4)	8.3 (7.0 - 9.8)	***Time**
	$bias$	-0.3 (-0.8 - 0.8)	-1.7 (-2.5 - -1.4)	***Time**
	CP_2	57.0 (45.5 - 65.0)	64.4 (49.3 - 69.4)	***Freq**
	iCP_5	13.3 (8.7 - 24.2)	16.5 (12.1 - 29.4)	
	TDI_{95}	8.4 (6.3 - 9.9)	11.2 (9.1 - 12.2)	***Time**
	$prop$	100.0 (98.0 - 100.0)	100.0 (96.8 - 100.0)	
ECG Fantasia	C	3.7 (2.3 - 6.0)	1.6 (1.1 - 3.7)	***Time**
	$2SD$	6.6 (5.8 - 7.2)	9.5 (7.8 - 10.2)	***Time**
	$bias$	-1.1 (-2.1 - 0.2)	-2.9 (-4.0 - -2.0)	***Time**
	CP_2	53.4 (46.9 - 62.1)	52.9 (43.9 - 69.2)	
	iCP_5	14.7 (10.1 - 20.3)	32.7 (18.7 - 40.5)	***Time**
	TDI_{95}	7.7 (6.9 - 8.6)	12.5 (10.7 - 13.2)	***Time**
	$prop$	100.0 (98.2 - 100.0)	100.0 (97.0 - 100.0)	
PPG RRest-healthy	C	1.3 (1.0 - 1.6)	0.7 (0.6 - 1.0)	***Time**
	$2SD$	9.4 (8.2 - 9.8)	11.9 (11.2 - 12.2)	***Time**
	$bias$	-2.3 (-2.7 - -0.3)	-5.5 (-5.7 - -5.0)	***Time**
	CP_2	37.2 (33.9 - 42.2)	35.7 (32.6 - 41.2)	
	iCP_5	29.8 (26.2 - 33.3)	48.0 (41.4 - 50.8)	***Time**
	TDI_{95}	11.4 (10.3 - 12.0)	16.4 (16.2 - 16.6)	***Time**
	$prop$	100.0 (93.2 - 100.0)	100.0 (97.0 - 100.0)	

significantly improved $2SD$ values, whilst producing either similar or worse CP_2 values. The underestimates have a large effect on $2SD$ as it is highly sensitive to outliers, but little effect on CP_2. Figure 6.7 also demonstrates that frequency-domain techniques very rarely result in overestimates of RR by more than 5 bpm.

Table 6.4 shows how the performance of algorithms varied in health according to whether or not a modulation fusion technique was used. Most performance statistics showed that the use

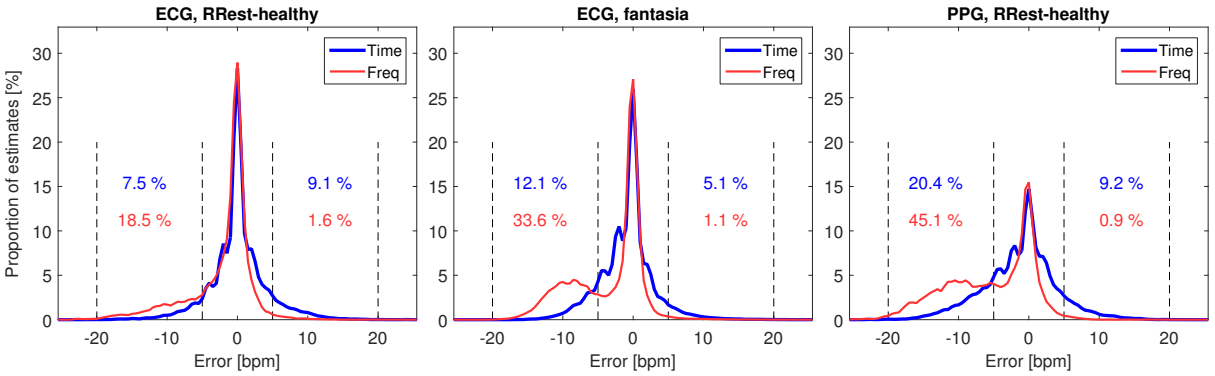

FIGURE 6.7: Comparison of RR algorithm errors when using time- or frequency-domain RR estimation techniques in health: The distributions of RR algorithm errors are shown when using algorithms containing time- and frequency-domain RR estimation techniques, applied to the ECG (left plot) and PPG (right).

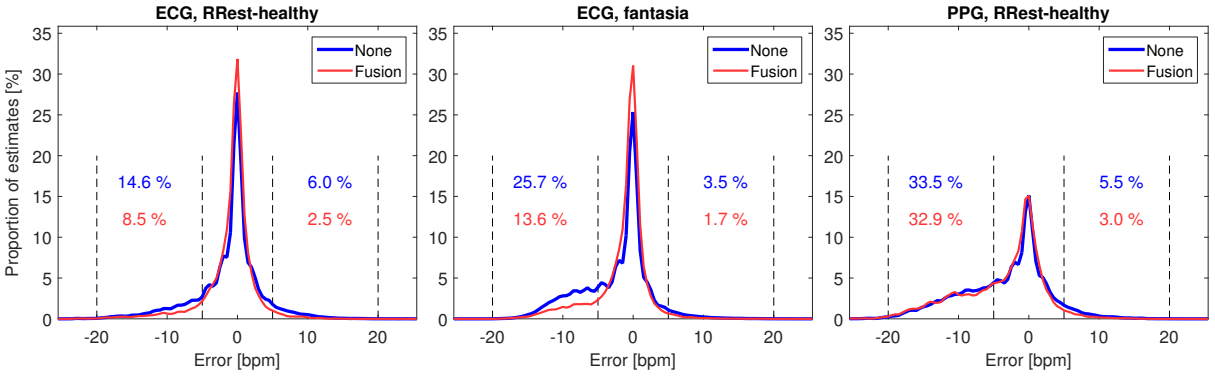

FIGURE 6.8: Comparison of RR algorithm errors when using or not using a modulation fusion technique in health: The distributions of RR algorithm errors are shown when using algorithms applied to the ECG (left and middle plots) and PPG (right).

of a modulation fusion technique significantly improved performance. The single exception was *prop*, which was significantly reduced when using a modulation fusion technique. This is to be expected, since modulation fusion techniques contain a quality assessment procedure resulting in RR not being estimated on some windows. Again, it is helpful to inspect how the distributions of RR algorithm errors are influenced by modulation fusion techniques, as shown in Figure 6.8. This figure demonstrates that algorithms using modulation fusion techniques were less vulnerable to underestimating RR when applied to the ECG (left and middle plots). In both these plots the proportion of estimates resulting in errors of between -5 and -20 bpm was much less when using modulation fusion techniques. However, modulation fusion techniques did not substantially reduce the vulnerability of algorithms to underestimating RR when applied to the PPG (right hand plot). This is shown by the similar proportions of underestimates regardless of whether or not a modulation fusion technique was used.

The performances of the highest ranked algorithms are provided in Table 6.5. Most of the

TABLE 6.4: Performances of algorithms with and without a modulation fusion technique in health. Significant differences are annotated by *, with the approach providing improved performance labelled.

		Algorithm performance, med (quartiles)		
Signal	Statistic	Fusion	Non-fusion	
ECG RRest-healthy	C	9.5 (4.5 - 11.8)	3.1 (1.7 - 4.9)	***Fusion**
	$2SD$	5.2 (4.9 - 6.6)	8.2 (7.0 - 8.9)	***Fusion**
	bias	-1.1 (-1.5 - -0.6)	-1.3 (-2.4 - -0.4)	
	CP_2	70.6 (65.6 - 72.6)	55.3 (46.4 - 64.8)	***Fusion**
	iCP_5	7.5 (6.4 - 13.7)	18.2 (13.1 - 28.6)	***Fusion**
	TDI_{95}	6.6 (5.7 - 9.2)	10.6 (8.5 - 12.0)	***Fusion**
	prop	82.0 (74.9 - 86.0)	100.0 (100.0 - 100.0)	***Non-fusion**
ECG Fantasia	C	4.76 (3.73 - 8.33)	1.88 (1.23 - 3.57)	***Fusion**
	$2SD$	6.1 (4.8 - 7.1)	7.9 (6.7 - 9.9)	***Fusion**
	bias	-1.9 (-2.1 - -1.2)	-2.6 (-3.9 - -1.1)	***Fusion**
	CP_2	68.7 (61.8 - 71.6)	50.6 (39.3 - 56.5)	***Fusion**
	iCP_5	14.2 (8.0 - 19.3)	27.3 (17.5 - 38.8)	***Fusion**
	TDI_{95}	10.0 (6.0 - 10.6)	10.7 (8.4 - 12.7)	***Fusion**
	prop	76.4 (63.0 - 86.7)	100.0 (100.0 - 100.0)	***Non-fusion**
PPG RRest-healthy	C	1.1 (0.8 - 1.5)	0.9 (0.7 - 1.2)	
	$2SD$	10.3 (8.1 - 11.0)	11.0 (9.6 - 12.1)	***Fusion**
	bias	-5.0 (-5.5 - -2.5)	-3.9 (-5.5 - -2.3)	
	CP_2	40.4 (38.0 - 43.9)	34.9 (32.1 - 40.7)	***Fusion**
	iCP_5	38.7 (28.3 - 45.3)	38.8 (31.1 - 49.3)	
	TDI_{95}	16.2 (11.0 - 16.5)	14.3 (11.7 - 16.4)	
	prop	77.6 (72.4 - 88.4)	100.0 (100.0 - 100.0)	***Non-fusion**

top-ranking algorithms used a time-domain RR estimation technique and a modulation fusion technique, as expected based on the earlier results. Algorithm $X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$ gave the best performance on both the ECG and PPG. It was consistently ranked most highly (according to the proposed cost function, C). Its LoAs were 0.2 ± 4.6 bpm and -0.1 ± 3.6 bpm when used with the ECG, and -0.3 ± 6.2 bpm when used with the PPG. This demonstrates good performance on the ECG, with 95% of errors expected to lie within approximately ± 5 bpm of the bias, and reasonable performance on the PPG, with a corresponding range of ± 7 bpm. These results are supported by the similar TDI_{95} values. Furthermore, when applied to the ECG the C values

TABLE 6.5: Performances of the ten highest ranked algorithms for the ECG and PPG in health: Ranked by C.

Signal	Algorithm	C	$2SD$ [bpm]	*bias* [bpm]	CP_2 [%]	iCP_5 [%]	TDI_{95} [bpm]	*prop* [%]
			RRest-healthy					
ECG	$X_{B1,2,3}E_{T4}F_{M1}$	15.7	4.6	0.2	72.2	4.6	4.8	82.0
	$X_{B1,2}E_{F3}F_{M5}$	12.8	4.8	-0.7	77.5	6.1	5.5	80.5
	$X_{B1,2}E_{T3}F_{M5}$	12.8	4.7	-0.3	72.0	5.6	5.1	86.5
	$X_{B1,2}E_{T2}F_{M5}$	12.5	4.7	-0.3	71.9	5.7	5.2	88.3
	$X_{B1,2}E_{F7}F_{M5}$	12.2	5.0	-1.0	77.7	6.3	6.1	78.4
	$X_{B1,2}E_{F6}F_{M5}$	12.0	4.8	-0.8	76.4	6.4	5.9	79.6
	$X_{B1,2,3}E_{T2}F_{M1}$	12.0	4.9	-0.7	67.5	5.7	5.4	85.1
	$X_{B1,2,3}E_{T3}F_{M1}$	10.4	5.0	-0.8	66.0	6.4	5.5	84.4
	$X_{B1,2,3}E_{F3}F_{M1}$	10.3	5.2	-1.1	74.4	7.2	6.3	62.1
	$X_{B1,2}E_{F1}F_{M5}$	9.9	5.5	-1.1	72.8	7.4	6.7	81.1
			Fantasia					
ECG	$X_{B1,2,3}E_{T4}F_{M1}$	36.6	3.6	-0.1	81.2	2.2	3.7	85.0
	$X_{B1,2}E_{T4}F_{M5}$	15.0	4.5	0.2	77.4	5.2	5.1	79.0
	$X_{B1,2,3}E_{T2}F_{M1}$	10.5	4.3	-1.3	62.9	6.0	5.3	86.9
	$X_{B1,2,3}E_{T3}F_{M1}$	9.5	4.2	-1.3	61.4	6.4	5.4	86.6
	$X_{B1,2,3}E_{T1}F_{M1}$	9.0	4.7	1.5	65.9	7.3	5.9	90.0
	$X_{B1,2}E_{T2}F_{M5}$	8.5	4.9	-0.9	68.1	8.0	5.9	75.6
	$X_{B1,2}E_{T3}F_{M5}$	8.3	4.9	-0.9	66.1	8.0	6.0	76.3
	$X_{B3}E_{T4}$	8.2	6.2	-0.2	71.4	8.8	6.8	98.0
	$X_{B5}E_{T4}$	7.0	6.3	0.2	68.2	9.7	7.0	99.1
	$X_{B3}E_{T1}$	6.8	6.7	1.4	65.1	9.6	7.5	100.0
			RRest-healthy					
PPG	$X_{B1,2,3}E_{T4}F_{M1}$	4.3	6.2	-0.3	54.1	12.5	7.1	72.5
	$X_{B3}E_{T4}$	2.8	8.1	0.3	52.3	18.5	8.9	96.7
	$X_{B1,2}E_{T4}F_{M5}$	2.8	7.1	-0.7	50.5	18.2	9.1	66.6
	$X_{B6}E_{T4}$	2.3	8.8	-0.4	48.4	21.3	9.9	97.5
	$X_{B1,2,3}E_{T2}F_{M1}$	1.8	7.7	-2.1	40.4	22.8	10.2	89.0
	$X_{B1,2,3}E_{T1}F_{M1}$	1.8	7.8	2.4	45.2	25.5	10.2	88.4
	$X_{B3}E_{T2}$	1.8	8.6	-2.0	44.5	25.2	10.4	100.0
	$X_{B3}E_{T3}$	1.7	8.6	-2.0	43.1	25.2	10.7	100.0
	$X_{B2}E_{T4}$	1.7	9.7	-0.3	43.4	25.3	10.9	97.3
	$X_{B1,2,3}E_{T3}F_{M1}$	1.7	7.8	-2.2	39.7	23.9	10.2	88.6

were 15.7 and 36.6, indicating a low proportion of erroneous RRs. However, when applied to the PPG the C value was 4.3, indicating that approximately one in four outputted RRs were erroneous. It should be noted that half of the top ten algorithms contained frequency-domain

RR estimation techniques when used with ECG signals in RRest-healthy, and these had higher CP_2 values than the remaining top-ten algorithms.

The algorithms performed significantly better when using the ECG, compared to the PPG. This was the case for all performance statistics except *prop*, which was significantly higher when using the PPG. The results of the comparison between using the ECG or PPG as the input signal are provided in Table A.6 in the appendix.

Algorithms tended to perform better on young subjects in the RRest-healthy Dataset, and on elderly subjects in the Fantasia dataset. The sizes of the differences between the groups were minimal, suggesting that the effect of age on performance is not clinically relevant. The results of the comparisons between age groups are provided in Table A.8 in the appendix.

6.3.5 Algorithm performances in hospitalised patients

The algorithms exhibited a wide range of performances in hospitalised patients, as demonstrated in Figure 6.9. The values of $2SD$ ranged from 2.5 bpm at best to 22.0 bpm at worst. Similarly, there were large variations in the other performance characteristics, with CP_2 ranging from 8.2 to 98.0 %, iCP_5 from 0.6 to 80.4 %, C from 0.1 to 159.5, and TDI_{95} from 0.74 to 28.6 bpm. The figure also shows variation between the datasets, with an apparent improvement in performance when algorithms were applied to the CapnoBase dataset compared to the MIMIC-II dataset.

Table 6.6 shows how the performance of algorithms varied according to whether a time- or frequency-domain RR estimation technique was used. Whereas in health time-domain RR estimation techniques consistently provided improved performance, there was not the case in hospitalised patients. Instead, frequency-domain techniques provided improved performance when algorithms were applied to ECG and PPG signals in the CapnoBase dataset; there was a moderate advantage to using frequency-domain techniques with ECG signals in the MIMIC-II dataset; and time-domain techniques provided improved performance on PPG signals in the MIMIC-II dataset. An explanation for the differences in performance was found by inspecting the differences in the distributions of RR algorithms errors, shown in Figure 6.10. This figure shows that time-domain techniques were vulnerable to overestimating RR on both datasets, and frequency-domain techniques were vulnerable to underestimating RR when used with MIMIC-II signals.

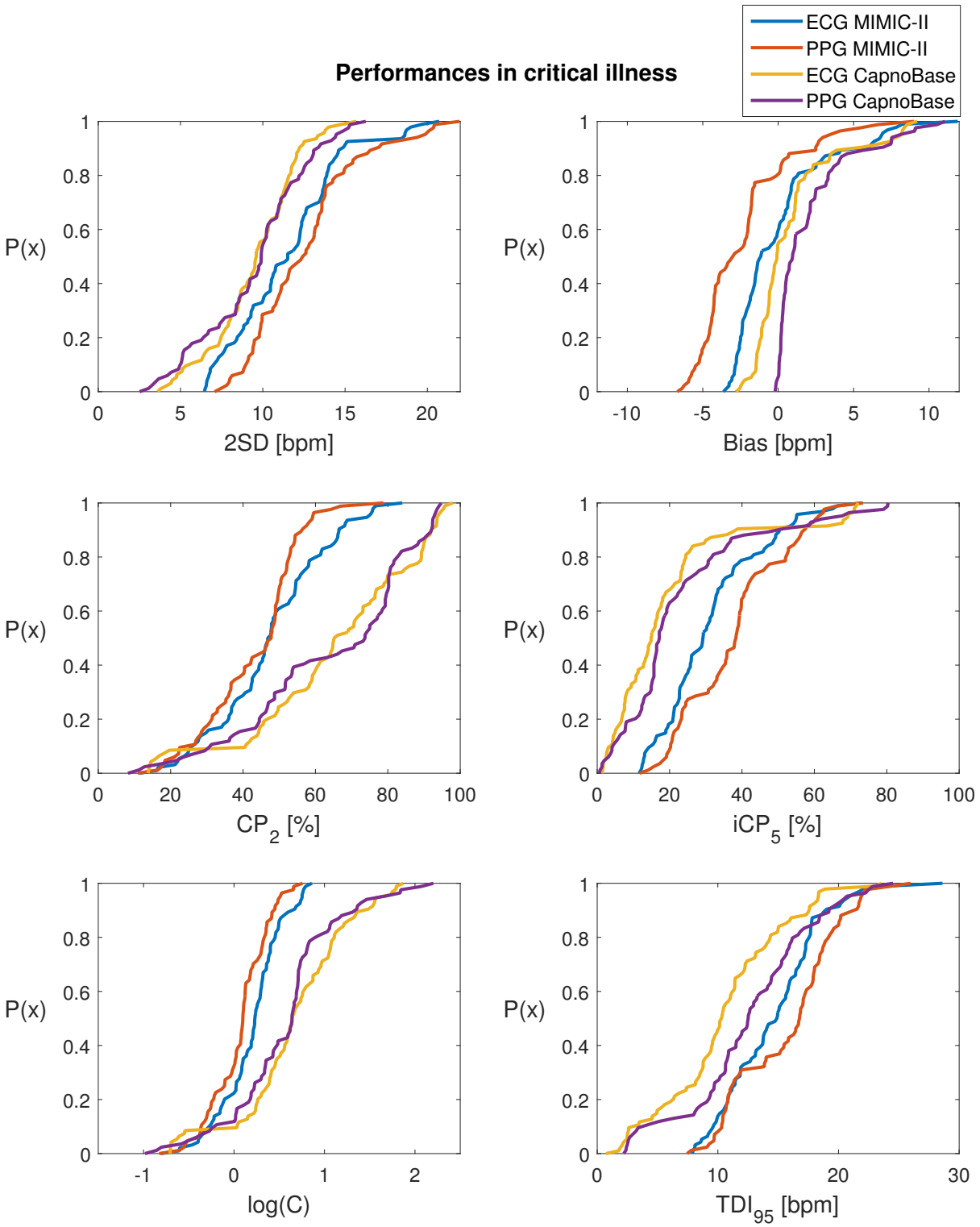

FIGURE 6.9: Performances of RR algorithms in hospitalised patients: Cumulative distribution functions of the performance characteristics of the algorithms assessed on the ECG and PPG signals in the CapnoBase and MIMIC-II datasets (see legend).

TABLE 6.6: Performances of algorithms using time- and frequency-domain RR estimation techniques in hospitalised patients. Significant differences are annotated by *, with the domain providing improved performance labelled.

		Algorithm performance, med (quartiles)		
Signal	Statistic	Time domain	Freq domain	
ECG CapnoBase	C	2.2 (1.1 - 4.1)	8.8 (4.5 - 21.0)	***Freq**
	$2SD$	9.4 (7.5 - 11.0)	10.2 (8.2 - 11.7)	
	bias	1.3 (1.0 - 3.6)	-1.0 (-1.4 - -0.4)	***Freq**
	CP_2	49.7 (41.6 - 61.1)	82.0 (70.6 - 90.4)	***Freq**
	iCP_5	23.5 (14.9 - 36.3)	9.4 (4.4 - 15.6)	***Freq**
	TDI_{95}	10.2 (8.6 - 14.2)	10.3 (4.6 - 12.4)	
	prop	98.7 (94.1 - 98.7)	98.7 (97.2 - 98.7)	
ECG MIMIC-II	C	1.4 (0.6 - 2.1)	2.0 (1.3 - 3.1)	***Freq**
	$2SD$	10.2 (8.7 - 12.4)	12.3 (10.5 - 13.9)	
	bias	0.9 (0.4 - 4.1)	-2.4 (-2.8 - -1.6)	
	CP_2	42.3 (27.7 - 46.8)	54.6 (46.2 - 66.1)	***Freq**
	iCP_5	29.3 (22.7 - 46.4)	29.0 (21.1 - 34.8)	
	TDI_{95}	11.9 (9.8 - 16.4)	15.7 (13.8 - 17.3)	***Time**
	prop	100.0 (99.0 - 100.0)	100.0 (94.3 - 100.0)	
PPG CapnoBase	C	1.6 (0.8 - 2.5)	5.3 (4.6 - 14.5)	***Freq**
	$2SD$	9.2 (7.5 - 11.5)	10.2 (6.9 - 11.6)	
	bias	3.2 (2.0 - 5.1)	0.3 (0.1 - 0.5)	***Freq**
	CP_2	46.4 (33.8 - 52.5)	80.4 (78.5 - 90.3)	***Freq**
	iCP_5	30.4 (21.6 - 43.9)	14.9 (6.3 - 16.4)	***Freq**
	TDI_{95}	12.0 (9.9 - 16.8)	12.7 (9.6 - 15.4)	
	prop	98.7 (85.5 - 98.7)	98.7 (88.1 - 98.7)	
PPG MIMIC-II	C	2.0 (1.0 - 2.6)	1.2 (0.6 - 1.3)	***Time**
	$2SD$	9.9 (9.1 - 11.6)	13.6 (12.6 - 14.6)	***Time**
	bias	-1.6 (-1.9 - 1.6)	-4.4 (-5.2 - -4.1)	***Time**
	CP_2	46.7 (35.2 - 53.6)	47.8 (32.6 - 50.4)	
	iCP_5	24.1 (20.7 - 35.0)	39.9 (38.4 - 52.9)	***Time**
	TDI_{95}	11.0 (10.4 - 14.6)	18.0 (16.9 - 20.0)	***Time**
	prop	98.3 (83.7 - 100.0)	100.0 (71.8 - 100.0)	

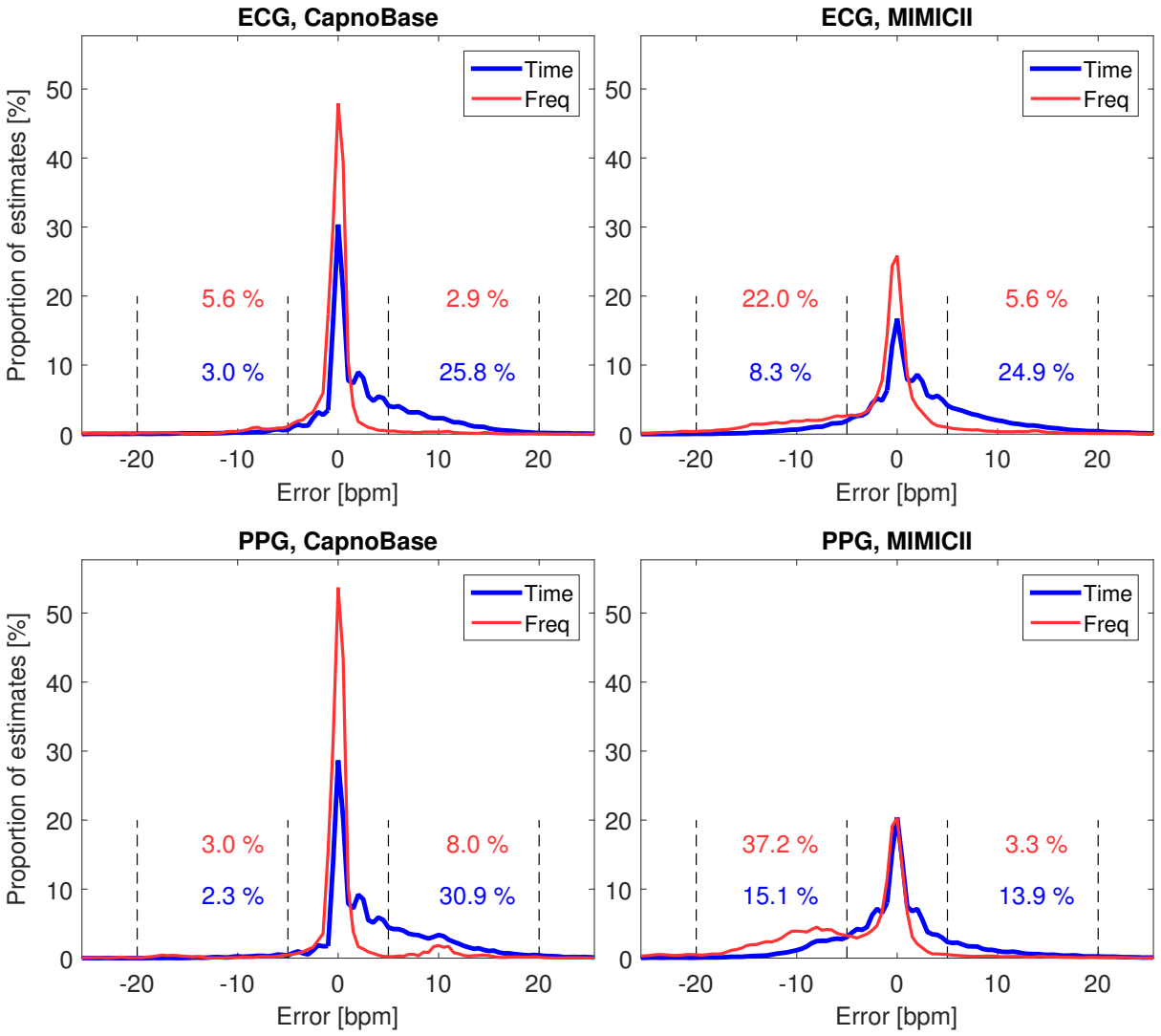

FIGURE 6.10: Comparison of RR algorithm errors when using time- or frequency-domain RR estimation techniques in hospitalised patients: The distributions of RR algorithm errors are shown when using algorithms containing time- and frequency-domain RR estimation techniques, applied to the ECG (upper plots) and PPG (lower plots).

Table 6.7 shows how the performance of algorithms varied in hospitalised patients according to whether or not a modulation fusion technique was used. Algorithms containing modulation fusion techniques showed improved performance when used on ECG signals from both datasets, and on PPG signals in CapnoBase. However, they showed no overall improvement when used with PPG signals on the MIMIC-II dataset. Again, it is helpful to inspect how the distributions of RR algorithm errors are influenced by using or not using a modulation fusion technique, as shown in Figure 6.11. This shows that modulation fusion techniques reduced the proportion of both under- and overestimates when using ECG and PPG signals in the CapnoBase dataset, and ECG signals in the MIMIC-II dataset. However, even when using modulation fusion techniques over 25 % of RR estimates derived from PPG signals in the MIMIC-II dataset remained at least

TABLE 6.7: Performances of algorithms with and without a modulation fusion technique in hospitalised patients. Significant differences are annotated by *, with the approach providing improved performance labelled.

		Algorithm performance, med (quartiles)		
Signal	Statistic	Fusion	Non-fusion	
ECG CapnoBase	C	21.0 (9.4 - 36.3)	3.8 (1.8 - 5.7)	***Fuse**
	$2SD$	6.4 (4.8 - 8.0)	10.5 (9.4 - 11.7)	***Fuse**
	$bias$	-0.3 (-0.6 - 1.1)	-0.0 (-1.2 - 1.4)	***Fuse**
	CP_2	89.0 (71.1 - 93.6)	61.2 (46.9 - 76.5)	***Fuse**
	iCP_5	4.4 (2.5 - 7.4)	17.0 (12.8 - 24.5)	***Fuse**
	TDI_{95}	4.6 (2.5 - 6.1)	11.4 (9.7 - 14.3)	***Fuse**
	$prop$	76.6 (72.0 - 80.2)	98.7 (98.7 - 98.7)	***None**
ECG MIMIC-II	C	3.3 (2.5 - 5.5)	1.6 (0.8 - 2.0)	***Fuse**
	$2SD$	7.3 (6.7 - 9.3)	12.3 (10.5 - 13.9)	***Fuse**
	$bias$	-1.4 (-1.6 - 0.7)	-0.8 (-2.5 - 1.0)	
	CP_2	56.5 (48.2 - 70.1)	44.4 (36.1 - 52.9)	***Fuse**
	iCP_5	16.2 (13.0 - 21.7)	31.9 (26.0 - 42.5)	***Fuse**
	TDI_{95}	11.4 (10.0 - 13.3)	15.7 (13.1 - 17.7)	***Fuse**
	$prop$	68.4 (50.7 - 74.5)	100.0 (100.0 - 100.0)	***None**
PPG CapnoBase	C	12.0 (2.8 - 26.5)	4.1 (1.5 - 5.1)	***Fuse**
	$2SD$	5.1 (4.4 - 7.3)	10.7 (9.7 - 12.5)	***Fuse**
	$bias$	0.2 (0.1 - 2.2)	1.1 (0.4 - 3.3)	***Fuse**
	CP_2	88.6 (52.8 - 92.4)	69.1 (46.4 - 79.4)	***Fuse**
	iCP_5	6.4 (3.5 - 16.4)	18.5 (15.7 - 31.8)	***Fuse**
	TDI_{95}	8.1 (2.9 - 9.7)	14.3 (11.7 - 16.4)	***Fuse**
	$prop$	69.4 (66.5 - 77.3)	98.7 (98.7 - 98.7)	***None**
PPG MIMIC-II	C	1.3 (1.0 - 2.4)	1.2 (0.6 - 1.8)	
	$2SD$	10.9 (8.8 - 11.9)	13.3 (10.0 - 14.9)	***Fuse**
	$bias$	-3.5 (-5.2 - -1.8)	-2.7 (-4.3 - -1.1)	
	CP_2	49.3 (39.9 - 54.7)	47.0 (31.5 - 50.0)	
	iCP_5	37.9 (20.6 - 41.4)	38.5 (26.1 - 52.4)	
	TDI_{95}	16.4 (10.3 - 19.2)	16.8 (11.3 - 18.5)	
	$prop$	59.8 (46.9 - 70.2)	100.0 (98.6 - 100.0)	***None**

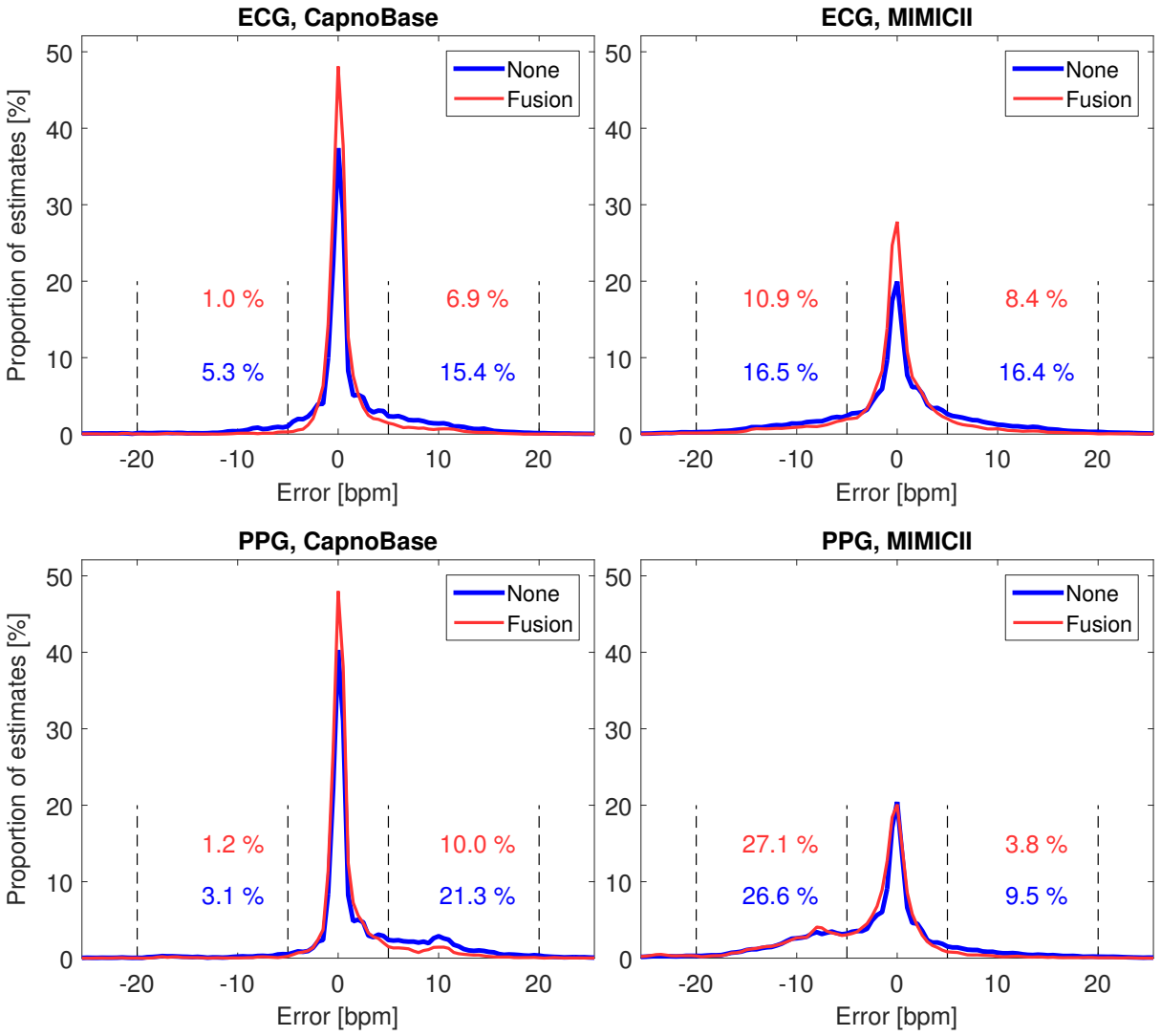

FIGURE 6.11: Comparison of RR algorithm errors when using or not using a modulation fusion technique in hospitalised patients: The distributions of RR algorithm errors are shown when using algorithms applied to the ECG (upper plots) and PPG (lower plots).

5 bpm below the reference RR.

There were some improvements in performance when using the ECG signal as the input, rather than the PPG, although these were not consistent across the datasets. There were no differences in $2SD$ or CP_2 values. These results are provided in full in Table A.7 in the appendix.

The performances of the highest ranked algorithms are provided in Table 6.8. All of the top-ranked algorithms applied to the ECG used frequency-domain RR estimation techniques. When using the PPG, all of the top-ranked algorithms on the CapnoBase dataset used frequency-domain RR estimation techniques, whereas all except one of those on the MIMIC-II dataset used time-domain techniques. Most of the top-ranked algorithms used a modulation fusion technique. $X_{\mathrm{B1,2,3}}E_{F7}F_{M2}$ provided high performance throughout. It achieved CP_2 values of

TABLE 6.8: Performances of the ten highest ranked algorithms for the ECG and PPG in hospitalised patients: Ranked by C.

Signal	Algorithm	C	$2SD$ [bpm]	$bias$ [bpm]	CP_2 [%]	iCP_5 [%]	TDI_{95} [bpm]	$prop$ [%]
ECG CapnoBase	$X_{B1,2,3}E_{F3}F_{M1}$	74.0	4.8	-0.3	91.8	1.2	2.5	75.9
	$X_{B1,2,3}E_{F1}F_{M1}$	62.2	6.4	-0.6	93.5	1.5	2.6	75.1
	$X_{B1,2,3}E_{F7}F_{M2}$	61.9	3.6	-0.1	98.0	1.6	0.7	83.2
	$X_{B1,2,3}E_{F6}F_{M1}$	55.7	6.3	-0.6	94.7	1.7	2.1	77.6
	$X_{B1,2,3}E_{F7}F_{M1}$	54.4	6.5	-0.6	94.1	1.7	2.2	76.3
	$X_{B1,2}E_{F6}F_{M5}$	37.0	8.1	-0.6	95.8	2.6	1.9	80.0
	$X_{B1,2}E_{F1}F_{M5}$	36.1	7.4	-0.6	91.7	2.5	2.6	81.5
	$X_{B1,2,3}E_{F2}F_{M1}$	35.4	5.8	-0.8	84.7	2.4	3.4	78.7
	$X_{B1,2,3}E_{F3}F_{M2}$	33.9	6.3	-0.3	95.4	2.8	1.8	93.6
	$X_{B1,2}E_{F3}F_{M5}$	30.4	7.8	-0.6	93.6	3.1	2.5	79.5
ECG MIMIC-II	$X_{B1,2,3}E_{F7}F_{M2}$	7.3	8.6	-1.2	84	11.6	12.1	49.4
	$X_{B1,2}E_{F6}F_{M5}$	6.2	7.3	-1.4	76.6	12.3	10.9	69.2
	$X_{B1,2}E_{F3}F_{M5}$	5.9	7.5	-1.3	75.2	12.7	10.7	71.3
	$X_{B1,2}E_{F7}F_{M5}$	5.8	7.6	-1.5	75.7	13.1	11.5	71.0
	$X_{B1,2,3}E_{F6}F_{M1}$	5.7	6.5	-1.5	68.3	12.1	10.5	50.8
	$X_{B1,2}E_{F1}F_{M5}$	5.7	7.8	-1.4	74.2	13.1	11.2	70.8
	$X_{B1,2,3}E_{F3}F_{M1}$	5.5	6.7	-1.4	68.0	12.4	11.0	49.8
	$X_{B1,2,3}E_{F7}F_{M1}$	5.4	6.8	-1.6	68.8	12.8	11.9	50.4
	$X_{B1,2,3}E_{F1}F_{M1}$	4.9	7.0	-1.5	66.4	13.6	11.4	49.4
	$X_{B1,2,3}E_{F3}F_{M2}$	4.7	9.3	-1.4	75.9	16.2	13.7	71.9
PPG CapnoBase	$X_{B1,2,3}E_{F1}F_{M1}$	159.5	3.1	-0.2	94.1	0.6	2.4	63.2
	$X_{B1,2,3}E_{F2}F_{M1}$	116.7	2.5	-0.2	92.8	0.8	2.4	70.3
	$X_{B1,2,3}E_{F6}F_{M1}$	69.8	3.3	0.0	93.6	1.3	2.2	69.6
	$X_{B1,2,3}E_{F7}F_{M1}$	68.4	3.5	-0.1	92.4	1.4	2.5	69.0
	$X_{B1,2,3}E_{F3}F_{M1}$	46.4	3.7	0.1	92.3	2.0	2.7	65.7
	$X_{B1,2}E_{F2}F_{M5}$	29.2	5.1	0.2	90	3.1	3.4	78.7
	$X_{B1,2}E_{F7}F_{M5}$	25.5	4.4	0.1	93.2	3.6	2.9	76.7
	$X_{B1,2,3}E_{F7}F_{M2}$	23.2	4.9	0.2	94.8	4.1	2.6	82.1
	$X_{B1,2}E_{F6}F_{M5}$	22.8	4.9	0.1	92.4	4.1	3.2	78.2
	$X_{B1,2}E_{F1}F_{M5}$	20.3	5.1	0.1	92.2	4.5	4.2	73.9
PPG MIMIC-II	$X_{B1,2}E_{T4}F_{M5}$	5.7	7.5	0.2	66.9	11.7	7.8	59.7
	$X_{B1,2,3}E_{F7}F_{M2}$	4.6	10	-1.8	78.8	17.3	14.0	39.6
	$X_{B1}E_{T4}$	4.6	8.9	0.1	64.3	14.1	9.1	97.3
	$X_{B1,2,3}E_{T4}F_{M1}$	3.4	7.1	0.7	52.1	15.5	7.5	39.8
	$X_{B6}E_{T4}$	3.3	9.9	0.7	59.1	18.2	10.1	97.1
	$X_{B5}E_{T4}$	3.1	10.0	0.3	59.3	19.3	10.4	97.8
	$X_{B4}E_{T4}$	3.0	9.9	0.2	58.2	19.1	10.6	96.9
	$X_{B1,2}E_{T5}F_{M5}$	2.9	7.8	-2.1	56.5	19.8	9.7	70.5
	$X_{B1,2}E_{T3}F_{M5}$	2.8	8.1	-1.8	57.6	20.3	9.7	70.0
	$X_{B1,2}E_{T2}F_{M5}$	2.6	9.1	-2.0	55.9	21.3	10.8	72.7

98.0 and 94.8 % when applied to CapnoBase ECG and PPG signals respectively, and 84.0 and 78.8 % when applied to MIMIC-II ECG and PPG signals. However, differences in performances between the two datasets make it difficult to predict how this and other algorithms would perform on other clinical datasets. Its LoAs when used with the ECG were -0.1 ± 3.6 bpm and -1.2 ± 8.6 bpm on the CapnoBase and MIMIC-II datasets respectively. Similarly, when used with PPG signals the LoAs were 0.2 ± 4.9 bpm and -1.8 ± 10.0 bpm respectively. Furthermore, both this algorithm and the other high-performing algorithms exhibited iCP_5 values of greater than 10 % when applied to MIMIC-II data, reflected in low C values of less than 10.

6.4 Discussion

This chapter presented a comprehensive assessment of the performances of RR algorithms in healthy and hospitalised patients. A total of 95 RR algorithms were assessed using the ECG as an input signal, and 85 using the PPG. The assessment was conducted on four publicly available datasets: two acquired from healthy subjects, and two from hospitalised patients. In contrast, previous studies of RR algorithms in the literature have assessed a maximum of 30 RR algorithms [278] using a maximum of two datasets (Section 4.4.4), with the exception of publications arising from this work [10]. This study provides a thorough assessment of existing RR algorithms in comparison. Furthermore, the assessment was conducted using the publicly available RRest Toolbox of algorithms [6, 10]. Consequently, it provides a foundation with which researchers can refine existing RR algorithms, and develop novel RR algorithms.

The performances of the RR algorithms varied greatly. Firstly, the set of 95 algorithms provided vastly different performances when assessed using any particular dataset and input signal. For instance, the smallest range of $2SD$ values exhibited by the algorithms when using any particular dataset and input signal was 6.17 - 12.45 bpm, when assessing the algorithms on the RRest-healthy Dataset using PPG signals. This indicates that the expected range of errors provided by the worst-performing algorithm can be twice that of the best-performing algorithm, demonstrating the importance of selecting the best-performing algorithm for clinical use. Secondly, the performances of algorithms varied across different datasets. For example, $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$ was considered one of the best algorithms in hospitalised patients. The reported $2SD$ values of this algorithm in hospitalised patients were 3.6 and 4.9 bpm when using ECG and PPG signals respectively from the CapnoBase dataset, compared to 8.6 and 10.0 bpm on the MIMIC-II dataset. This demonstrates the importance of dataset selection for algorithm assessment. In

this instance, it is notable that most of the subjects in the CapnoBase dataset were younger than acutely-ill patients who might be monitored using a wearable sensor with an inbuilt RR algorithm. The CapnoBase dataset contains data from 29 paediatrics and 13 adults. Furthermore, the adults were relatively young, aged a median (lower - upper quartiles) of 47.0 (39.0 - 54.5) years. They were much younger than the LISTEN Trial subjects (Chapter 3) who were aged 67.5 (60.0 - 76.0) years. Therefore, results from the CapnoBase dataset, whilst useful, may not be representative of those that could be expected in acutely-ill adult patients. In contrast, the MIMIC-II dataset contains only data from adults staying in critical care, who may be more representative of acutely-ill adult patients than those in CapnoBase.

Key features of RR algorithms were identified which provide improved performance. In health, improved performance was achieved by using a time-domain RR estimation technique, and by using a modulation fusion technique. Time-domain RR estimation techniques provided improved performance since they did not under-estimate RRs as frequently as frequency-domain techniques. This effect was greatest with the PPG signal, where 45.1 % of RR estimates derived using frequency-domain techniques underestimated the reference RR by between 5 and 20 bpm. Modulation fusion techniques provided improved performance when using ECG signals as they reduced the proportion of underestimated RRs. However, modulation fusion techniques did not substantially reduce the proportion of underestimated RRs when using PPG signals. In contrast, frequency-domain techniques provided improved performance in hospitalised patients, apart from when using PPG signals from the MIMIC-II dataset, where algorithms containing time-domain techniques were superior. Similarly, modulation fusion techniques remained beneficial, providing improved performance apart from when using PPG signals from the MIMIC-II dataset. Once again, modulation fusion techniques were not able to reduce the large proportion of underestimated RRs when using the PPG signals in the MIMIC-II dataset. The choice of input signal also impacted performance, with ECG signals providing improved performance in health, and on the MIMIC-II dataset.

The best-performing algorithms were $X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$ in health, and $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$ in hospitalised patients. The performances of these algorithms are summarised in Table 6.9. For comparison, the LoAs of the routinely used ImP RR monitoring technique have been previously reported as -0.2 $\pm$ 5.4 bpm [10], or worse [32]. Therefore, one could assume that an algorithm with similar or better performance would be suitable for use in wearable sensors. Based on this, $X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$ appears to perform sufficiently well when used with the ECG and healthy subjects. However, it showed poorer performance in hospitalised patients, suggesting that it

TABLE 6.9: A summary of the performances of the best RR algorithms.

Dataset	$X_{B1,2,3}E_{T4}F_{M1}$		$X_{B1,2,3}E_{F7}F_{M2}$	
	ECG	PPG	ECG	PPG
		$LoAs, bias \pm 2SD$ [bpm]		
		Healthy subjects		
RRest-healthy	0.2 ± 4.6	-0.3 ± 6.2	-1.8 ± 7.9	-5.0 ± 10.2
Fantasia	-0.1 ± 3.6	n/a	-1.8 ± 7.8	n/a
		Hospitalised patients		
CapnoBase	2.2 ± 5.3	3.2 ± 5.6	-0.1 ± 3.6	0.2 ± 4.9
MIMIC-II	2.5 ± 7.1	0.7 ± 7.1	-1.2 ± 8.6	-1.8 ± 10.0
		Cost function, C		
		Healthy subjects		
RRest-healthy	15.7	4.3	6.1	1.4
Fantasia	36.6	n/a	5.3	n/a
		Hospitalised patients		
CapnoBase	3.6	2.0	61.9	23.2
MIMIC-II	2.0	3.4	7.3	4.6

may not be suitable for use with acutely-ill subjects. $X_{B1,2,3}E_{F7}F_{M2}$ showed poorer performance when assessed using the LoAs technique, with the exception of the CapnoBase dataset results, which cannot be considered representative of acutely-ill patients for the reasons detailed earlier. Therefore, it appears that none of the algorithms assessed are suitable for use with acutely-ill patients. The C statistic is also informative, indicating how many more times likely it is that an RR estimate will have an error of < 2 bpm, than an error of > 5 bpm. The C values were relatively low ($C < 10$, indicating that more than one in ten RR estimates had absolute errors of > 5 bpm) with the exceptions of: (i) $X_{B1,2,3}E_{T4}F_{M1}$ applied to the ECG in health; and (ii) $X_{B1,2,3}E_{F7}F_{M2}$ applied to the CapnoBase data. None of the algorithms reached $C > 10$ when assessed on the MIMIC-II dataset. This confirms that although $X_{B1,2,3}E_{T4}F_{M1}$ may perform sufficiently well with the ECG in health, none of the algorithms appear to be suitable for continuous monitoring of acutely-ill patients.

The issues with the best-performing algorithms are further illustrated in Figure 6.12. The issues identified with time- and frequency-domain RR estimation techniques remain in the best-performing algorithms, despite them both containing a modulation fusion technique. The figure shows that the frequency-domain algorithm ($X_{B1,2,3}E_{F7}F_{M2}$) frequently underestimated RR by

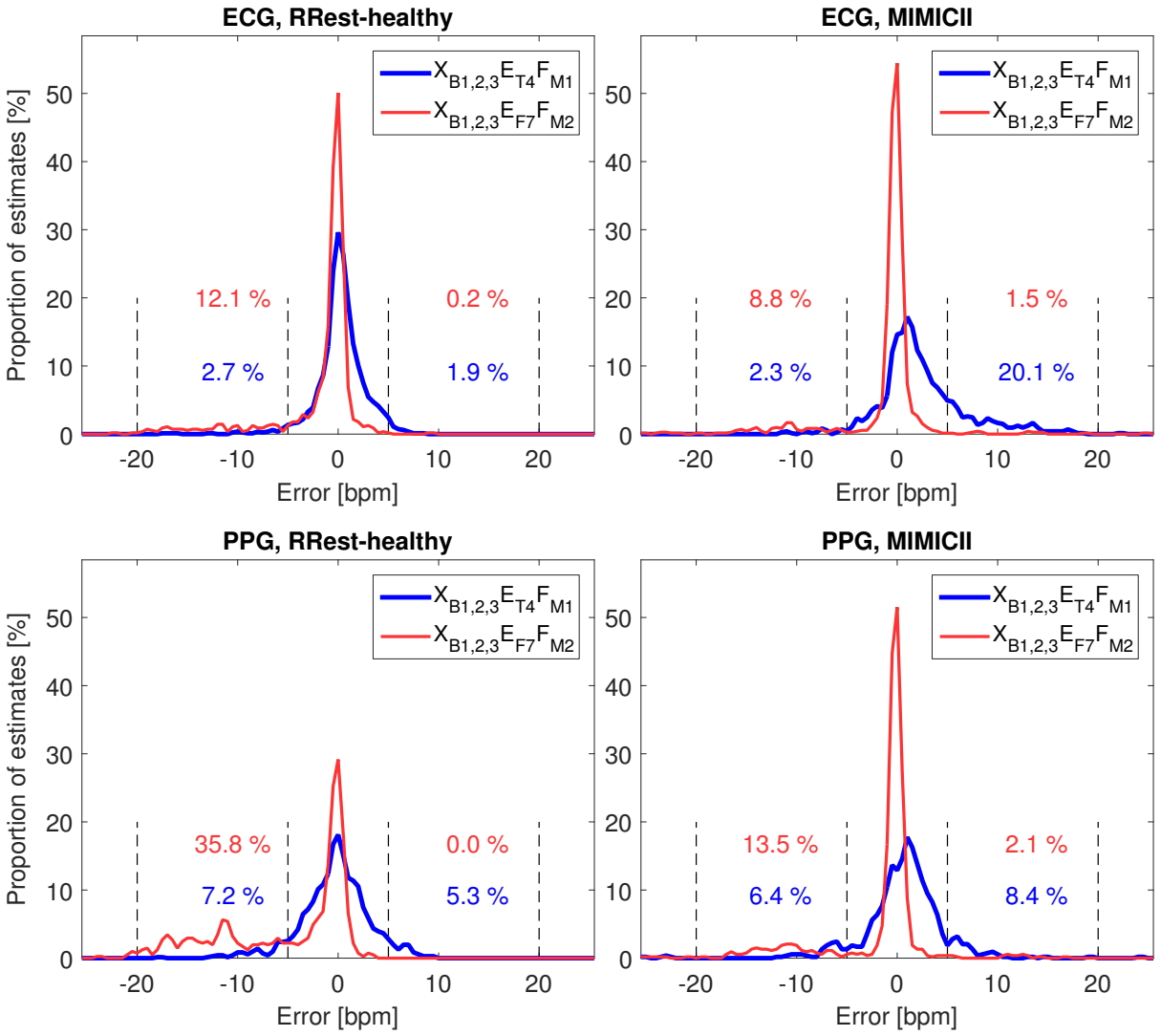

FIGURE 6.12: The distributions of RR algorithm errors for the best algorithms used with hospitalised patients: The distributions of RR algorithm errors are shown when using $X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$, which contains a time-domain RR estimation technique, and $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$, which contains a frequency-domain RR estimation technique.

more than 5 bpm. The time-domain algorithm ($X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$) showed a wider spread of errors, indicating a lower level of precision. In addition, on the MIMIC-II dataset it frequently overestimated RR.

Therefore, this assessment suggests that a novel algorithm is required to estimate RR from the ECG or PPG for use in wearable sensors. One approach to improve algorithms which has recently been proposed is to incorporate an algorithm stage in which the quality of the extracted respiratory signals is assessed [154, 155]. The techniques which could be used at this stage, known as respiratory quality indices (RQIs), are similar to SQIs in that they assess the quality of a continuous signal, and are used to eliminate low quality signals. However, RQIs

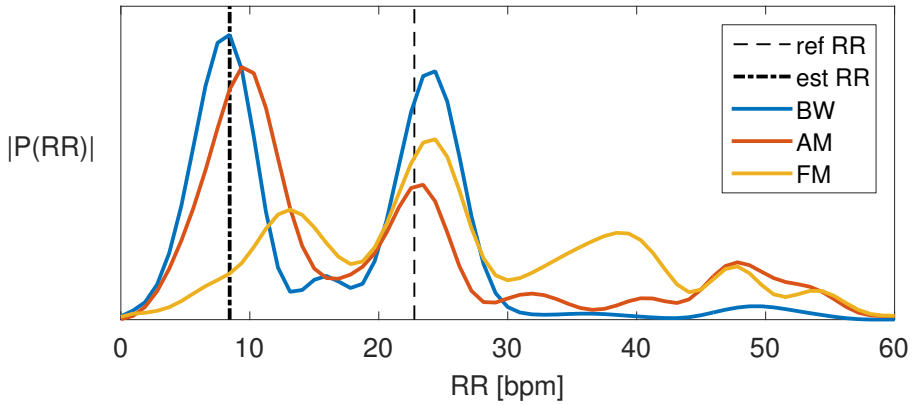

FIGURE 6.13: The rationale for a respiratory quality index (RQI): This shows the power spectra of the three input signals fused by $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$ for a particular window. The RR was substantially underestimated due to the algorithm detecting non-respiratory, low frequency content, despite the true RR being manifested in all three power spectra. It may be possible to develop an RQI to mitigate against such errors.

assess whether or not sufficient respiratory modulation is present in a signal, in contrast to SQIs which assess whether a signal contains high-fidelity cardiac modulation. Inspection of windows on which RR algorithms failed suggests that RQIs may improve RR algorithm performance. For instance, Figure 6.13 shows the power spectra of the three input signals fused by $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$ for a particular window. In this example all three power spectra showed peaks at the reference RR. However, the RR was underestimated because two of the power spectra showed larger peaks at a much lower RR. A RQI could potentially be used to eliminate any input signals which exhibit multiple peaks in their power spectra, thereby avoiding this issue.

In this study a novel cost function was proposed to rank RR algorithms. The cost function was calculated primarily using CP_2 and iCP_5, and indicates the relative likelihood of an RR estimate being within $\pm$ 2 bpm of the reference RR, compared to being outside $\pm$ 5 bpm of the reference. This approach was chosen to reflect the potential utility of RR algorithms for use with wearable sensors. The rationale for its design is based on the potential use of an RR algorithm with the national early warning score (NEWS, Section 2.2). The NEWS assigns a score to RR (and other physiological parameters) based on its deviation from normal values. When calculating the NEWS (Table 2.2) an error in RR of 2 bpm can only change a patient's calculated risk of deterioration by one level out of four. In contrast, an error of 5 bpm can change the calculated risk of deterioration from no risk up to the highest level (and vice versa). Therefore, the CP_2 and iCP_5 statistics were chosen to reflect the potential for either false detection of deteriorations, or missing deteriorations, based on erroneous RR values. Other statistics were also reported as together they provide a more comprehensive description of algorithm performance.

6.5 Final Remarks

This study assessed the performance of algorithms to estimate RR from the ECG and PPG in healthy and hospitalised patients. A total of 95 algorithms were assessed using four datasets. The main finding was that $X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$ performed sufficiently well for use with the ECG in healthy subjects. However, no algorithms were found to perform sufficiently well with hospitalised patients. The best-performing algorithm in hospitalised patients was $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$, although this did not perform sufficiently well with data from critical care. Therefore, further refinement of algorithms is required before their use in wearable sensors to monitor acutely-ill patients.

Key features of algorithms were identified which provide improved performance. Firstly, the use of a modulation fusion technique to fuse RRs derived from multiple respiratory signals improved performance when using ECG signals. Secondly, time-domain RR estimation techniques provided improved performance in healthy subjects, and frequency-domain RR estimation techniques provided superior performance when used with ECG signals in hospitalised patients. Thirdly, algorithms performed better in health when using ECG signals rather than PPG signals.

The RRest-healthy dataset, code to download and appropriately format the Fantasia, CapnoBase and MIMIC-II datasets, RRest Toolbox of algorithms, and analysis code used in this study are being made publicly available at `http://peterhcharlton.github.io/RRest`. These resources allow future researchers to reproduce the analyses presented here, and to develop and evaluate novel RR algorithms.

Chapter 7

An Algorithm for Continuous Respiratory Rate Monitoring using the Electrocardiogram and Photoplethysmogram

This chapter presents a novel algorithm for continuous monitoring of respiratory rate (RR) in hospitalised patients using the electrocardiogram (ECG) and photoplethysmogram (PPG). The algorithm was designed to harness the strengths of both time- and frequency-domain RR estimation techniques. The algorithm was optimised using data from healthy subjects, and assessed using data from hospitalised patients. Its performance was compared to that of a control algorithm, the best performing algorithm in Chapter 6. In addition, analyses were conducted to determine whether performance varied between subjects and at different RRs. The results were used to evaluate the potential utility of the algorithm for use with the ECG and PPG, and to identify particular areas for further clinical testing.

7.1 Introduction

Existing RR algorithms have been designed for use in clinical settings where it is important that a RR measurement is provided for most, or all, windows of input signal. This design approach is appropriate in settings such as exercise stress tests [345], where RR changes rapidly, or for

making diagnoses in an outpatient clinic [80], where a single measurement is required in a short space of time. In these settings there is a high penalty for not outputting a measurement at any particular time, since this will interrupt real-time monitoring, or prolong the time taken to make a diagnosis in busy clinics. However, the downside to this design approach is that less importance is placed on ensuring that the estimated RRs are as precise as possible, which results in frequent erroneous RRs (see Section 6.4).

In contrast, when monitoring acutely-ill patients for detection of deteriorations there is no need to output a RR for most windows of input signal. This is because the physiological changes which accompany acute deteriorations occur over the course of hours [18–22], so one can afford to analyse several minutes of input signal before outputting a RR. However, in this setting there is a high penalty for outputting an erroneous RR since it could be mistaken for an early sign of deterioration requiring a clinical response. False alerts of deteriorations waste clinical resources, since a clinician typically has to visit a patient showing signs of deterioration. In addition, they may desensitise clinicians to true alerts, a phenomenon known as alarm fatigue [117], reducing the effectiveness of clinical responses when they are truly needed.

More recently some RR algorithms have been developed without the requirement of outputting a RR for every window of input signal [80, 109, 147, 154]. These algorithms only output a RR at times when the estimated RR is judged to be of high quality. This is achieved by either (i) analysing the quality of the extracted respiratory signal from which an RR is estimated [109, 154], or (ii) assessing the similarity of multiple RRs generated simultaneously [80] or within a short time period [147]. Indeed, techniques such as Smart Fusion [80], which incorporates a quality assessment step to determine whether to output the current estimate, have been shown to improve the precision of RR algorithms at the expense of outputting RRs less frequently [10]. Nonetheless, even algorithms which are not constrained to outputting RRs for every window of input signal can have a high error rate (see Section 6.4), with over 10 % of RR estimates having errors of > 5 bpm (breaths per minute). An error of this magnitude is enough to raise a false alert for a patient showing no signs of deterioration (as explained in Section 6.4). Therefore, further research is required to design an algorithm specifically for continuous monitoring of acutely-ill patients.

This chapter presents the design and assessment of a novel RR algorithm specifically designed for continuous monitoring of hospitalised patients to detect deteriorations. The rationale for the design is presented in Section 7.2. Section 7.3 describes the methods used to optimise the

design of the algorithm, and to assess its performance, including a comparison of its performance with that of the best-performing algorithm identified in Chapter 6 (the control algorithm). The results are described in Section 7.4. The implications of the results, and the potential utility of the novel algorithm for use in wearable sensors are discussed in Section 7.5.

7.2 Algorithm Design

RR algorithms for use with the ECG and PPG broadly consist of three fundamental stages [4] (see Figure 6.1): extraction of a respiratory signal, estimation of RR, and fusion of multiple RR estimates. Several studies have compared techniques for extracting respiratory signals [157, 159, 179, 186, 268, 330], and techniques for fusion of RR estimates have recently been the subject of much research [4, 80, 113, 237, 238, 341]. However, there has been relatively little research into the relative performances of algorithms using different RR estimation techniques [109]. The novel algorithm presented here exploits the differing performance characteristics of different RR estimation techniques to improve the precision of RR estimation.

RR estimation techniques act in either the time- or the frequency-domain (detailed in Section 4.4.3.3). Time-domain techniques identify individual breaths within a signal, and then estimate RR from the time intervals between each breath. In contrast, frequency-domain techniques transform the signal into a power spectrum, and the RR is estimated as the frequency corresponding to the maximum power spectral density. Techniques from both domains have been widely used in the literature (see Table 4.7).

Time- and frequency-domain RR estimation techniques have differing performance characteristics. This is demonstrated in Figure 7.1, which shows plots illustrating the performances of time- and frequency-domain techniques from both the literature and novel analyses. The plots corresponding to time-domain techniques (upper panel of four plots) exhibit a wide spread of estimates around the line of identity, indicating moderate precision. In contrast, frequency-domain techniques exhibit a smaller spread of estimates around the line of identity indicating high precision, but also have frequent erroneous estimates resulting in estimated RRs of (approximately) < 12 bpm. Consequently, neither approach is ideal for use with wearable sensors: time-domain techniques are not highly precise; and frequency-domain techniques, whilst highly precise for much of the time, are also vulnerable to large errors.

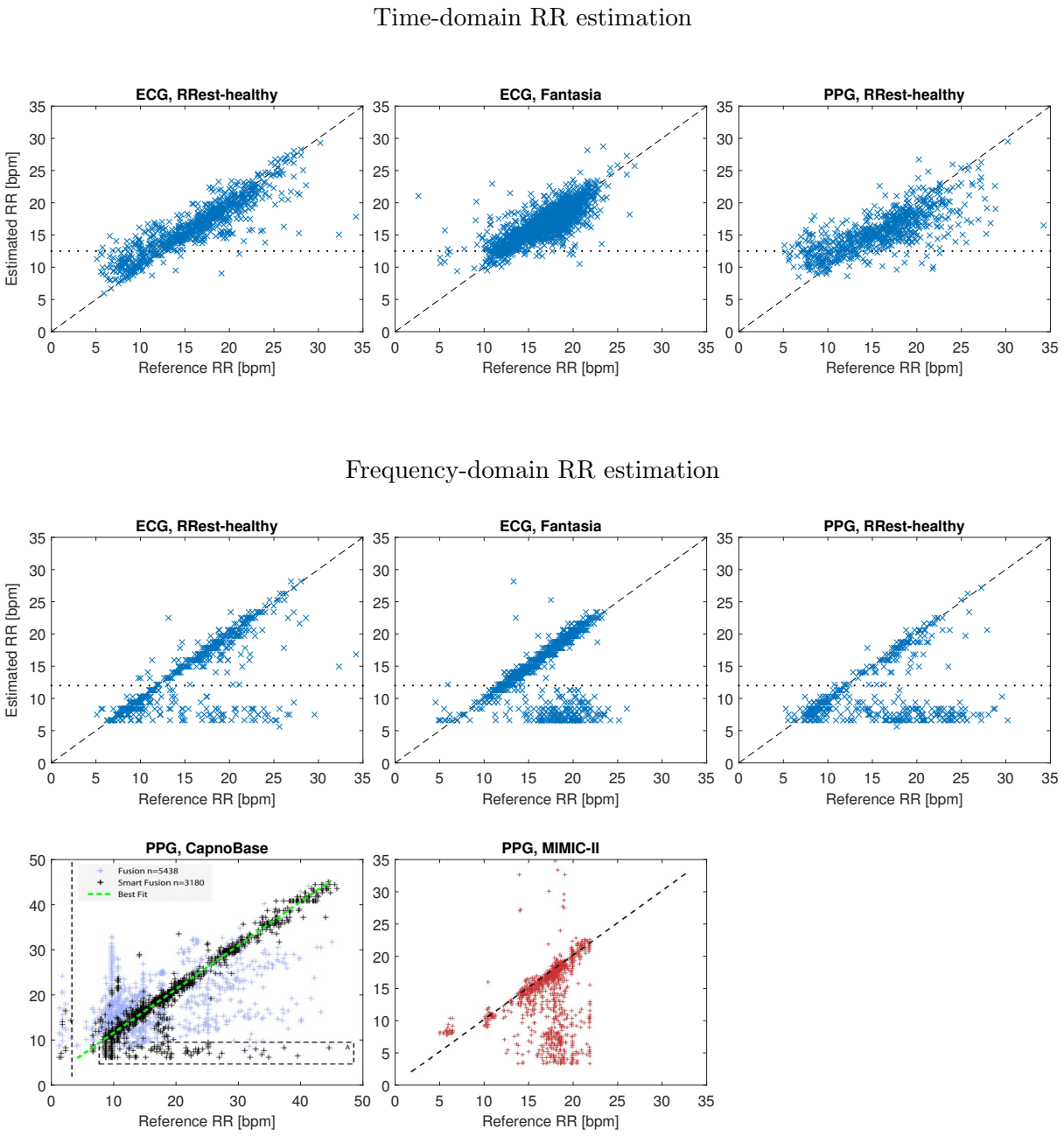

FIGURE 7.1: Scatter plots of RR estimates produced by RR algorithms containing time- and frequency-domain RR estimation techniques: The upper panel of three plots shows estimates produced by algorithms containing time-domain techniques, and the lower panel of five plots shows estimates produced by frequency-domain techniques. The upper panel and the upper row of the lower panel show plots produced from analyses of publicly available datasets (using the best-performing algorithms from Chapter 6: $X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$ and $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$). The lower row of the lower panel shows plots adapted from the literature: PPG, CapnoBase [80] © 2013 IEEE, and PPG, MIMIC-II [341] © 2016 IEEE. The frequency-domain plots show several underestimates lying below the line of identity, occurring at estimated RRs of (approximately) < 12 bpm. This characteristic is used in the novel RR algorithm to identify and mitigate against underestimates produced by frequency-domain techniques.

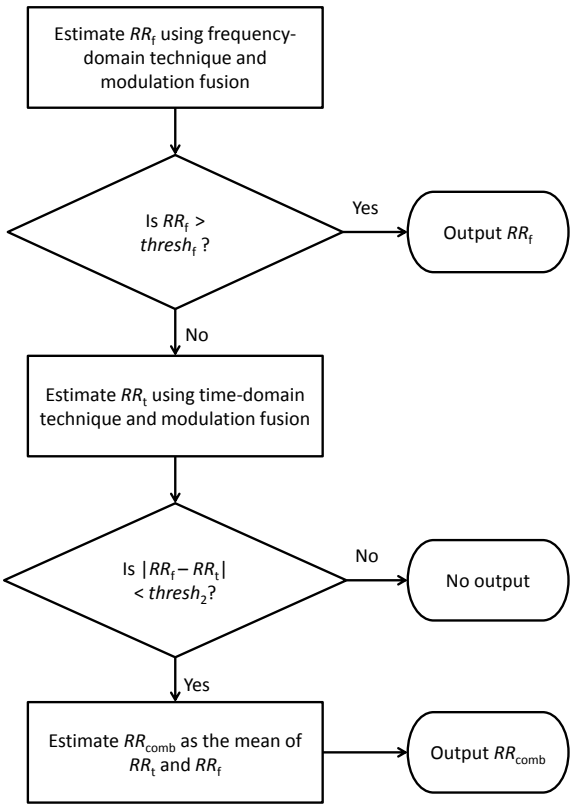

FIGURE 7.2: The novel RR algorithm: The algorithm primarily uses a frequency-domain technique to estimate RR (RR_{f}), although if the estimated RR is $< thresh_{\mathrm{f}}$ then a time-domain technique is also used. If the difference between the time-domain estimate (RR_{t}) and RR_{f} is less than $thresh_{\mathrm{comb}}$ then a combined estimate (RR_{comb}) is outputted. Otherwise, no RR is outputted.

The novel RR algorithm presented in this chapter was designed to combine the strengths of time- and frequency-domain RR estimation techniques. The key feature of the algorithm is its ability to use the high-precision RR estimates produced by frequency-domain RR estimation techniques, whilst identifying and mitigating against those with large errors. Underestimates produced by frequency-domain techniques share a common feature which facilitates their identification and exclusion: they manifest at estimated RRs of (approximately) $<$ 12 bpm (see bottom row, Figure 7.1). This was, in part, reported in Chapter 6, where it was observed that algorithms containing frequency-domain RR estimation techniques frequently underestimate RR, but very rarely overestimate RR (see Figure 6.8). The underestimates produced by frequency-domain techniques have also been noted in [80, 277, 341]. In contrast, time-domain techniques do not tend to produce large underestimates (top panel, Figure 7.1). This difference in performance characteristics is used by the novel algorithm to identify and mitigate against underestimates produced by frequency domain techniques.

The novel RR algorithm is summarised in Figure 7.2. Firstly, an algorithm containing a

frequency-domain RR estimation technique was used to produce a candidate RR. This was performed using either $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M1}}$ or $X_{\mathrm{B1,2}}E_{\mathrm{F7}}F_{\mathrm{M5}}$ (both algorithms were trialled). The frequency-domain RR estimation technique, E_{F7}, was chosen based on its high performance in Chapter 6 (for definitions see Tables 4.5, 4.7, and 4.8). The threshold value used for quality assessment during RR estimation was either $thresh_2$ or $thresh_3$ depending on whether two or three inputs were used (F_{M1} uses three, whereas F_{M5} uses only two). If the RR estimate produced was above a threshold RR, $thresh_{\mathrm{f}}$, then it was accepted as the final RR estimate since it was likely to be highly precise. If the RR estimate was below this threshold, then a second RR estimate was obtained using an algorithm containing a time-domain technique. This was performed using either $X_{\mathrm{B1,2,3}}E_{\mathrm{T4}}F_{\mathrm{M1}}$ or $X_{\mathrm{B1,2}}E_{\mathrm{T4}}F_{\mathrm{M5}}$. The time- and frequency-domain RR estimates were then fused using F_{M5}. In this work F_{M5} was performed by calculating the difference between the two estimates, and outputting the mean of the two estimates if the difference was below $thresh_2$. If the difference was greater than $thresh_2$ then no RR estimate was outputted.

The following stages of the algorithm were optimised during the design process:

- The threshold RR, $thresh_{\mathrm{f}}$, above which frequency-domain estimates were accepted required optimisation. The bottom row of plots in Figure 7.1 suggested that a value of approximately 12 bpm may be appropriate.
- The quality assessment threshold used for the fusion of three RR estimates in F_{M1}, $thresh_3$, was optimised. It was expected that a value of approximately 3 - 5 bpm would be appropriate, in line with the value of 4 bpm proposed in [80].
- The quality assessment threshold used for the fusion of two RR estimates in F_{M5}, $thresh_2$, was optimised. It was expected that a value of approximately 3 bpm would be appropriate, based on previous work (Section 6.3.3).

Each of these stages was optimised in the study now presented to produce an algorithm suitable for continuously monitoring of acutely-ill patients.

7.3 Methods

7.3.1 Datasets

Four publicly available datasets were used in this study: RRest-healthy [7], Fantasia [302, 303], CapnoBase [80] and MIMIC-II [299, 303]. Since these datasets were also used for the study presented in Chapter 6, they are only briefly summarised here (for further details see Section 6.2.1). The datasets contain simultaneous ECG, PPG, and respiratory signals (except Fantasia, which does not contain PPG signals). CapnoBase also contains manual breath annotations. The RRest-healthy and Fantasia datasets were acquired from young and elderly healthy subjects, whereas CapnoBase and MIMIC-II were acquired from hospitalised patients. The datasets acquired from healthy subjects were used for algorithm optimisation, and those acquired from hospitalised subjects were used for assessing the algorithm's performance.

7.3.2 RR estimation

RR estimation was performed using the RRest Toolbox of algorithms, which is a publicly-available framework with which to assess the performance of both existing and novel RR algorithms [6, 10] (described in Section 4.6.3). Since the processing steps performed by the toolbox were described in detail in Section 6.2, they are only briefly recapped here. Firstly, the quality of each 32 s window of ECG, PPG and reference respiratory signals was assessed. The signal quality index (SQI) algorithms described in Section 3.3 were used for the ECG and PPG. The quality of each of the reference respiratory signals was assessed using the algorithms described in Section 6.2.2, including the novel SQI described in Section 4.6.2. Any windows in which any of the required signals were of low quality were excluded from the analysis. Secondly, RRs were estimated from the ECG and PPG using two algorithms: the novel RR algorithm, and the algorithm previously identified as the most suitable existing algorithm for continuous monitoring of hospitalised patients (Section 6.4): $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$ (definitions provided in Tables 4.5, 4.7 and 4.8). Thirdly, reference RRs were estimated from the manual breath annotations in the CapnoBase dataset, and from the reference respiratory signals in the other datasets (see Section 6.2.5 for details of the algorithms used).

7.3.3 Statistical analysis

The performance of each algorithm was assessed in the same manner as for the study in Chapter 6. The performances of algorithms were quantified using the following statistics (as detailed in Section 6.2.6):

- *bias*: The systematic bias, indicating an algorithm's accuracy.
- $2SD$: The expected range encompassing 95% of errors around the *bias*.
- CP_2: The proportion of RR estimates which have an absolute error of < 2 bpm.
- iCP_5: The proportion of RR estimates which have an absolute error of > 5 bpm.
- TDI_{95}: The 95^{th} percentile of absolute errors, indicating the absolute error below which 95 % of absolute errors fall.
- *prop*: The proportion of windows for which an algorithm provides a RR estimate, rather than no output.
- C: A bespoke cost-function indicating the clinical utility of an algorithm: CP_2/iCP_5 .

Algorithms were ranked according to C, since this statistic was designed to assess the clinical utility of algorithms for continuous monitoring of acutely-ill patients. Further details of the rationale for C can be found in Sections 6.2.6 and 6.4.

7.3.4 Algorithm optimisation

The optimal configurations of the novel RR algorithm were identified by assessing the algorithm's performance when using a range of configurations. The following options were trialled:

- $thresh_f = \{5.0, 5.5, 6.0, ..., 20.0\}$ bpm
- $thresh_2 = \{0.5, 1.0, 1.5, ..., 8.0\}$ bpm
- $thresh_3 = \{0.5, 1.0, 1.5, ..., 8.0\}$ bpm

This resulted in a total of 7936 configurations when using F_{M1} (which takes three input signals and therefore uses $thresh_3$), and 496 configurations when using F_{M5} (which takes two input

signals and therefore does not use $thresh_3$). The novel algorithm's performance was optimised using the two datasets acquired from healthy subjects: RRest-healthy and Fantasia. For this process the two datasets were each split according to young and elderly subjects, resulting in four data subsets. Optimal configurations were identified for eight design specifications formed from combinations of: using the ECG or PPG signal; using either two or three input RR estimates (using either X_{B1} and X_{B2}, or X_{B1}, X_{B2} and X_{B3}); and insisting on a *prop* of at least either 20 % or 40 %. The C and *prop* statistics were used for this process, since ideally an algorithm would have both a high C and a high *prop* value. The performances of the configurations were ranked according to their C values on each data subset. Any configurations which had less than the required *prop* on any data subset were excluded. The lowest ranking of each configuration across the data subsets was used as its overall ranking. The optimal configuration for each design specification was then the one with the highest overall ranking.

The configurations of the novel algorithm were denoted as (for example):

$$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 20\%) \quad ,$$

indicating the respiratory signals used (X_{B1}, X_{B2} and X_{B3}), the RR estimation techniques used (E_{F7} and E_{T4}), the modulation fusion method (F_{M1}), and the design configuration (E for ECG, and a minimum *prop* of 20 %).

The performances of the novel algorithm were assessed when using each of the eight identified configurations. This was performed on all four datasets, although particular importance was placed on the performances on the MIMIC-II dataset, since the patients in this dataset were most similar to those of acutely-ill hospitalised patients, such as those in the LISTEN Trial (described in Chapter 3). Performance was assessed using the C and *prop* statistics. The most suitable configurations for use with the ECG and PPG were selected from the four configurations for each signal.

7.3.5 Performance assessment

The performance of the novel algorithm was compared to that of a control algorithm: $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$, which has previously been identified as the most suitable existing algorithm for continuous monitoring of hospitalised patients (Section 6.4). This comparison was performed to

determine whether the novel algorithm would be more suitable for use in this clinical scenario than the best existing algorithm.

Firstly, the performances were compared using the statistics described in Section 7.3.3. This was performed using the datasets acquired from hospitalised patients: CapnoBase and MIMIC-II. In addition, the MIMIC-II dataset was used to analyse the performances of the two algorithms across individual subjects, and across a range of reference RRs. These latter analyses were used to determine whether the algorithm performed consistently across different subjects and RRs. These were important since respiratory signals extracted from the ECG and PPG are of lower quality at higher RRs (see Section 5.5), and the performance of RR algorithms can vary greatly between different subjects [278]. Only subjects which contributed at least 5 high quality windows were included in the analyses across individual subjects to avoid spurious results. The C and *prop* statistics were used for the analysis across RRs. The CP_2 and *prop* statistics were used for the analysis between subjects since often sufficient windows were not available to calculate C robustly for individual subjects.

7.4 Results

7.4.1 Algorithm optimisation

Figure 7.3 shows an example of the procedure used to optimise the novel algorithm. This figure demonstrates the selection of optimal configurations for each of the following design specifications: $X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 20\%)$ and $X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 40\%)$, indicated by diamond and circular red dots respectively on ECG plots; $X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(P, 20\%)$ and $X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(P, 40\%)$, indicated by diamond and circular red dots respectively on PPG plots. These optimal configurations did not necessarily provide the best performance on any particular dataset. For instance, $X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 20\%)$ never provided the best performance on any of the individual data subsets. However, it gave the best overall performance across the four data subsets according to the ranking criteria used. The choice of algorithm configuration had a large impact on the algorithm's performance, with a large range of C and *prop* values produced by different configurations.

The optimal configurations of the novel algorithm identified for each design specification are listed in Table 7.1. The values of $thresh_2$ and $thresh_3$ tended to be lower at smaller specified

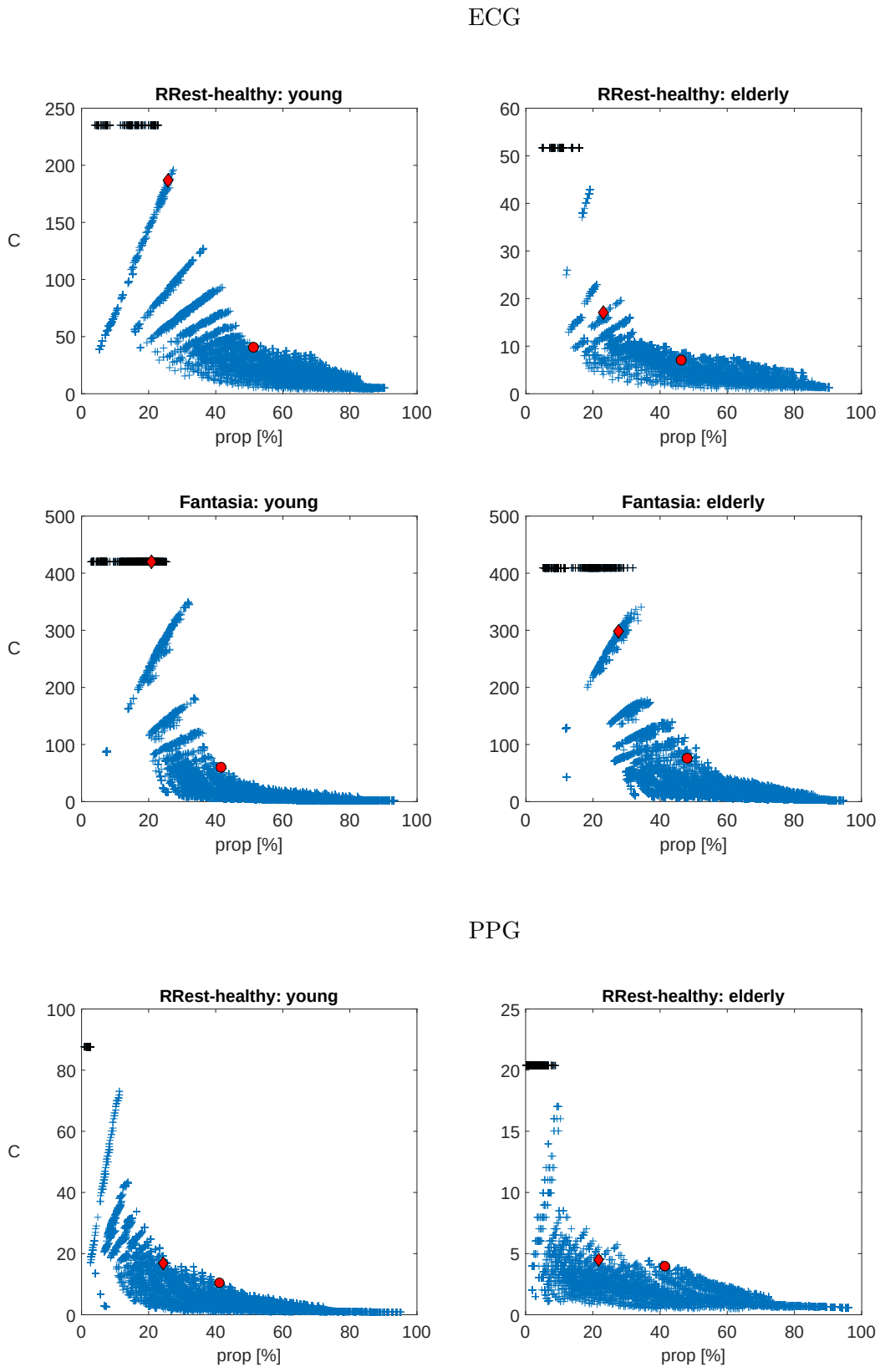

FIGURE 7.3: The performances of the novel RR algorithm across a range of configurations: Each plot shows the performances of the 7936 assessed configurations of the novel algorithm, quantified using the cost function, C, and the proportion of windows for which a RR estimate was provided, *prop*. Each individual configuration is shown by a blue cross. The performances of the optimal configurations for the ECG and PPG are shown in red, with a diamond red dot indicating a *prop* of $\geq$ 20 %, and a circular red dot indicating a *prop* of $\geq$ 40 %. Some configurations had infinite C values (when the denominator of C, iCP_5, was zero), which have been re-scaled and are shown in black. Note that y-axis scales vary between plots.

TABLE 7.1: Optimal configurations of the novel RR algorithm for different design specifications

Abbr.	Design specification			Configuration		
	Signal	No. inputs	Specified *prop* [%]	$thresh_f$ [bpm]	$thresh_2$ [bpm]	$thresh_3$ [bpm]
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 20\%)$	ECG	2	20	18.5	1.5	n/a
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 40\%)$	ECG	2	40	18.0	3.0	n/a
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 20\%)$	ECG	3	20	18.0	1.5	2.0
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 40\%)$	ECG	3	40	18.0	3.5	5.0
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(P, 20\%)$	PPG	2	20	16.0	3.5	n/a
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(P, 40\%)$	PPG	2	40	15.0	6.0	n/a
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(P, 20\%)$	PPG	3	20	12.0	5.0	2.5
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(P, 40\%)$	PPG	3	40	15.5	5.0	5.0

TABLE 7.2: The performances of the optimised configurations of the novel RR algorithm

Algorithm	RRest-healthy		Fantasia		CapnoBase		MIMIC-II	
	C	*prop*	C	*prop*	C	*prop*	C	*prop*
ECG								
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 20\%)$	31.7	40.1	160.0	28.9	64.5	25.0	28.7	31.7
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 40\%)$	28.3	55.8	44.3	41.8	63.0	36.5	26.4	40.3
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 20\%)$	59.5	25.2	543.0	24.1	∞	18.5	18.7	17.3
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(E, 40\%)$	20.6	50.0	67.2	44.7	59.3	35.6	15.0	29.6
PPG								
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(P, 20\%)$	6.8	23.0	n/a	n/a	17.5	38.8	7.9	28.6
$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(P, 40\%)$	2.8	45.5	n/a	n/a	20.2	53.2	5.1	40.3
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(P, 20\%)$	11.5	23.7	n/a	n/a	203.0	39.0	5.5	19.3
$X_{\mathrm{B1,2,3}}E_{\mathrm{F7,T4}}F_{\mathrm{M1}}(P, 40\%)$	7.8	41.1	n/a	n/a	70.3	45.9	4.3	23.6

prop values. This is because lower thresholds result in fewer RR estimates being deemed to be of high quality. In addition, $thresh_f$ was lower in configurations optimised for the PPG (approximately 15 bpm) than for the ECG (approximately 18 bpm).

The performances of the optimal configurations of the novel algorithm on each dataset are shown in Table 7.2. Performances are reported for all four datasets, although the performances

TABLE 7.3: Performances of the novel algorithm, $X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(\ldots)$, and the control algorithm, $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$ (identified in Chapter 6 as being most suitable for continuous monitoring of hospitalised patients): Ranked by C.

Signal & Dataset	Algorithm	C	$2SD$ [bpm]	*bias* [bpm]	CP_2 [%]	iCP_5 [%]	TDI_{95} [bpm]	*prop* [%]
ECG CapnoBase	$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 40\%)$	63.0	1.5	0.4	97.4	1.5	1.5	36.5
	$X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$	61.9	3.6	-0.1	98.0	1.6	0.7	83.2
ECG MIMIC-II	$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 40\%)$	26.4	3.2	0.0	85.8	3.2	4.5	40.3
	$X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$	7.3	8.6	-1.2	84	11.6	12.1	49.4
PPG CapnoBase	$X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$	23.2	4.9	0.2	94.8	4.1	2.6	82.1
	$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(P, 20\%)$	17.5	3.3	0.8	92.8	5.3	6.9	38.8
PPG MIMIC-II	$X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(P, 20\%)$	7.9	9.0	-0.2	82.0	10.4	9.4	28.6
	$X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$	4.6	10.0	-1.8	78.8	17.3	14.0	39.6

on the MIMIC-II dataset (shaded) were used to select the final configurations for use with the ECG and the PPG. On this dataset, the use of two input signals (F_{M5}) rather than three (F_{M1}) always resulted in higher C and *prop* values. Therefore, only configurations containing F_{M5} were considered. $X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 40\%)$ was chosen for use with the ECG signal over $X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(E, 20\%)$, since it provided RR estimates with a similar C value (26.4 compared to 28.7), for a greater proportion of windows (40.3 compared to 31.7 %). For the PPG, $X_{\mathrm{B1,2}}E_{\mathrm{F7,T4}}F_{\mathrm{M5}}(P, 20\%)$ was chosen since it had the highest C value. This was chosen irrespective of *prop* since none of the PPG configurations provided a particularly high C value. Only the chosen configurations were carried forward to the performance assessment now presented.

7.4.2 Performance assessment

Table 7.3 provides a comparison of the performances of the novel algorithm and the most suitable existing algorithm for continuous monitoring of hospitalised patients (the control algorithm, $X_{\mathrm{B1,2,3}}E_{\mathrm{F7}}F_{\mathrm{M2}}$, identified in Chapter 6). On the CapnoBase dataset there was no advantage to using the novel algorithm, since it provided similar C values to the control algorithm, but produced RR estimates for approximately half the proportion of windows compared to the control algorithm. In contrast, on the MIMIC-II dataset the novel algorithm provided improved C values for both the ECG and PPG, at the expense of only a small reduction in *prop*. Indeed, when using the ECG signal the novel algorithm had a C value of 26.4, returning RR estimates

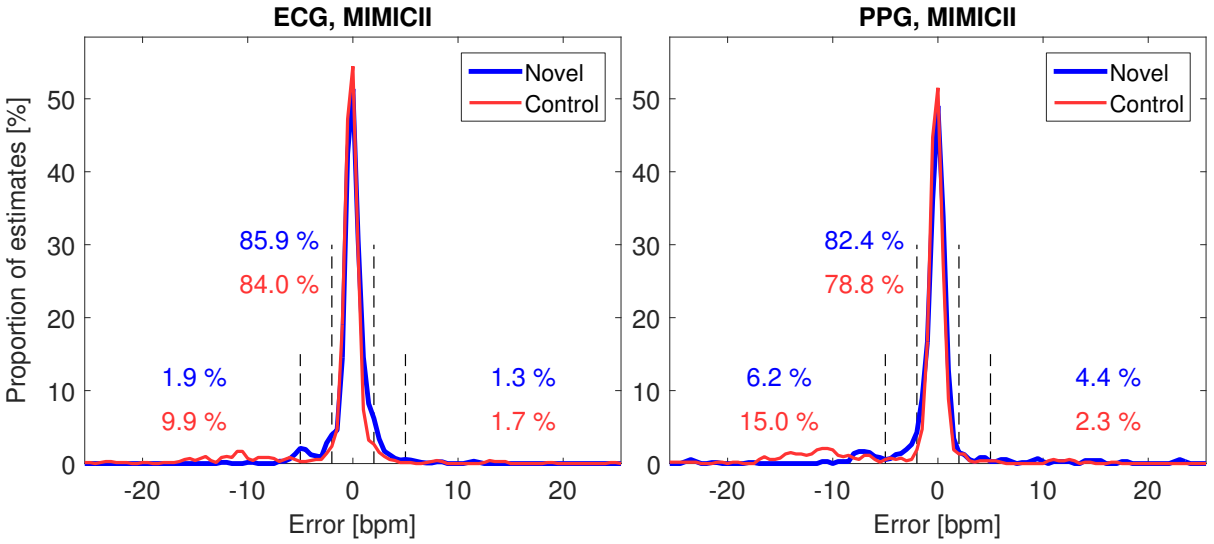

FIGURE 7.4: The error distributions of the novel and control RR algorithms when used with the ECG (left hand plot) and PPG (right hand plot). Results are from the MIMIC-II dataset.

for 40.3 % of windows, compared to $C = 7.3$ for the control algorithm across 49.4 % of windows. Furthermore, the *LoAs* improved from -1.2 $\pm$ 8.6 bpm to 0.0 $\pm$ 3.2 bpm when using the novel algorithm. Only 3.2 % of the RR estimates provided by the novel algorithm had absolute errors of > 5 bpm. There was some benefit to using the novel algorithm with the PPG, since the C value improved from 4.6 to 7.9. However, the *LoAs* remained similarly large between the two algorithms: -0.2 $\pm$ 9.0 bpm for the novel algorithm, compared to -1.8 $\pm$ 10.0 bpm for the control algorithm.

The distributions of the errors produced by the novel algorithm and the control algorithm are shown in Figure 7.4. The novel algorithm produced fewer underestimates (errors $<$ -5 bpm) than the control algorithm (a reduction from 9.9 to 1.9 % with the ECG, and 15.0 to 6.2 % with the PPG). This suggests that the fusion of time- and frequency-domain RR estimates was effective in mitigating against underestimates. The scatter plots in Figure 7.5 demonstrate the smaller number of underestimates produced when using the novel algorithm. However, despite the improvement, some large underestimates remained when using the novel algorithm with the PPG.

The performances of the novel algorithm and the control algorithm on individual subjects are summarised in Figure 7.6. This figure shows the distributions of *prop* and CP_2 values calculated for each individual subject. Both algorithms exhibited wide variations in *prop* and CP_2 values between subjects, regardless of the input signal. This suggests that the performance of the algorithms may vary between subjects when used in clinical practice.

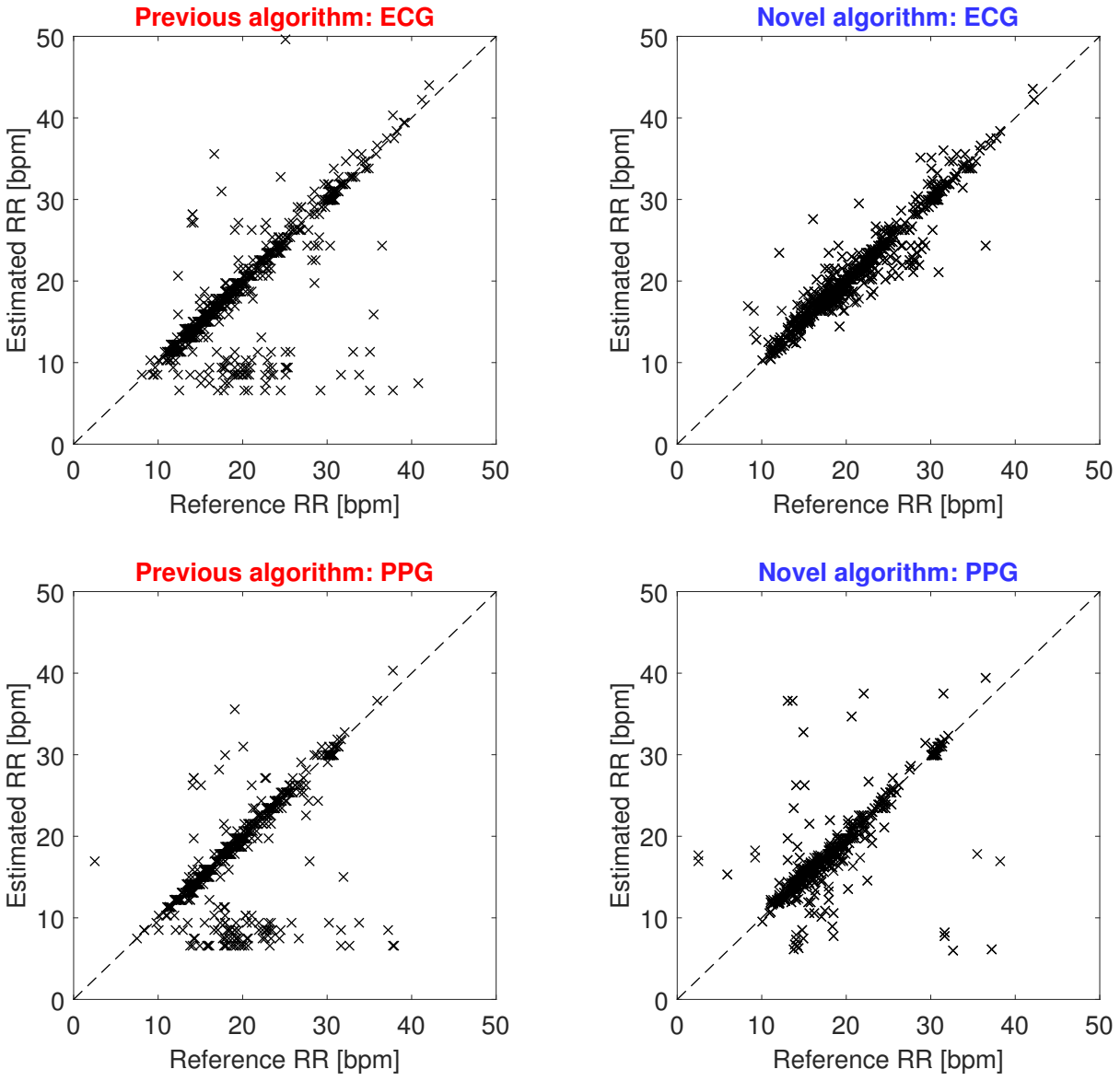

FIGURE 7.5: Scatter plots of RR estimates produced by the novel and the control RR algorithms: Results are shown for the ECG (upper plots) and PPG (lower plots). Results are from the MIMIC-II dataset.

The performances of the novel algorithm and the control algorithm at different reference RRs are shown in Figure 7.7. When using the ECG, both algorithms had lower C values at elevated RRs. In addition, the novel algorithm produced RR estimates for a smaller proportion of windows at lower RRs than at higher RRs. When using the novel algorithm with the PPG the variation in performance at different RRs was smaller, although the algorithm had lower C values at extreme RR values than at normal RR values. It also produced RR estimates for a higher proportion of windows at lower RRs than at higher RRs, in contrast with the ECG.

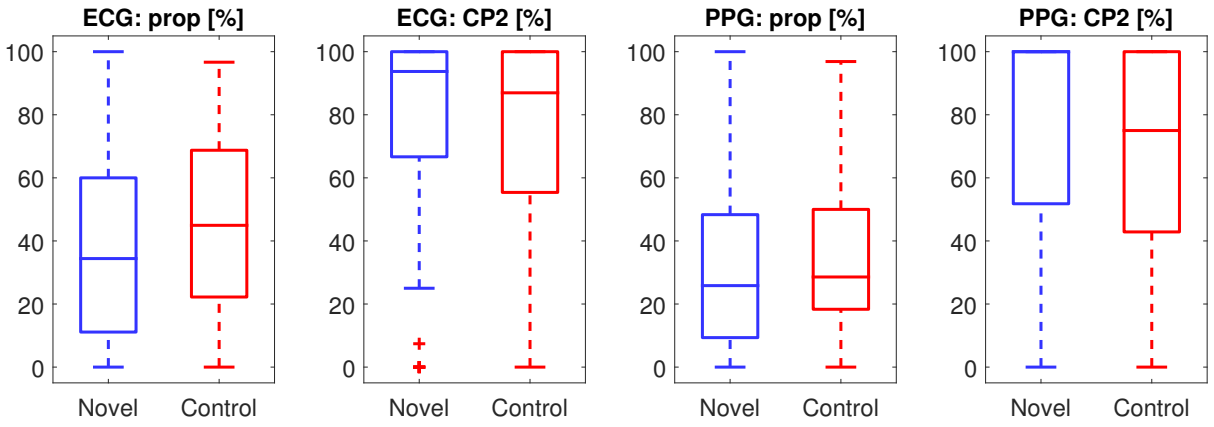

FIGURE 7.6: The performances of the novel and the control RR algorithms on individual subjects: The proportion of windows for which an RR estimate was returned (*prop*), and the proportion of RR estimates within 2 bpm of the reference RR (CP_2), are shown for the two algorithms when using the ECG (first and second plots) and PPG (third and fourth plots). Results are from the MIMIC-II dataset.

7.5 Discussion

In this chapter a novel algorithm was presented to estimate RR from the ECG and PPG, designed specifically for continuous monitoring of acutely-ill hospitalised patients. The algorithm uses both time- and frequency-domain RR estimation techniques, whereas previously proposed RR algorithms tend to use a single RR estimation technique. The algorithm was optimised using data from healthy subjects, and assessed on data from hospitalised patients. It performed well when used with the ECG, giving *LoAs* (within which 95 % of errors are expected to lie) of 0.0 ± 3.2 bpm and providing RR estimates for 40.3 % of windows. This level of performance is suitable for clinical use, particularly as only 3.2 % of estimates had absolute errors of > 5 bpm. Indeed, it may allow RR to be monitored unobtrusively using wearable sensors equipped with ECG monitoring. However, both the novel algorithm and the control algorithm did not perform well enough for continuous monitoring when used with the PPG.

The novel algorithm consists of the fusion of time- and frequency-domain RR estimates derived from baseline wander (BW) and amplitude modulation (AM) signals extracted from either the ECG or PPG signal. The addition of RR estimates derived from frequency modulation (FM) signals was detrimental to performance, so FM-based estimates were not included in the chosen configurations. This may be because the quality of FM signals diminishes with age (see Section 5.5). It has been observed both previously and in the current work that frequency-domain RR estimation techniques have a tendency to underestimate RRs, perhaps due to the presence of low frequency oscillations such as Traube-Hering-Mayer waves [80, 277]. The novel algorithm

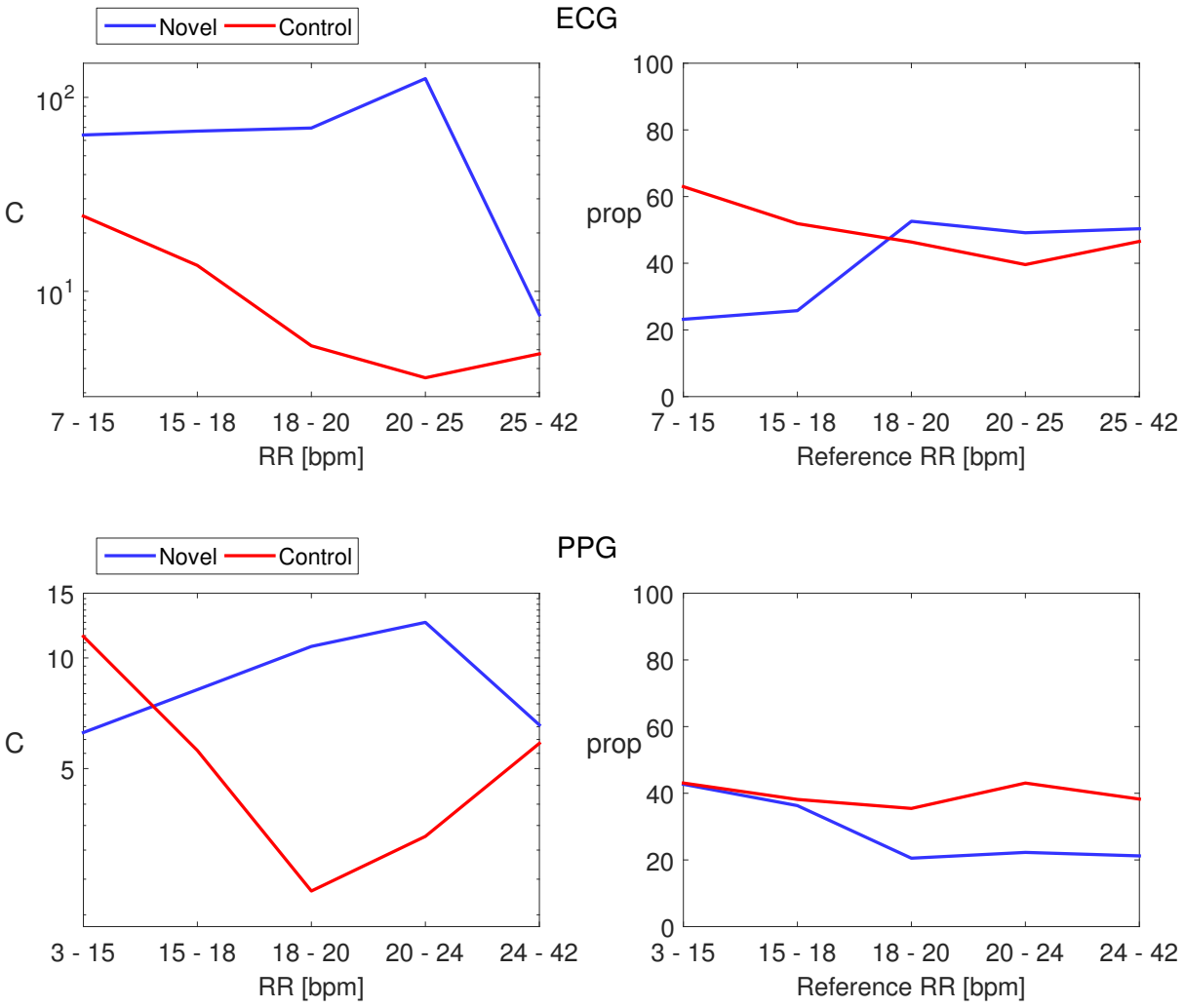

FIGURE 7.7: The performances of the novel and the control RR algorithms at different reference RRs: The bespoke cost-function (C, indicating the clinical utility of an algorithm), and the proportion of windows for which an RR estimate was returned (*prop*), are shown for the five quintiles of reference RRs. Performances are shown for the ECG (upper plots) and PPG (lower plots). Results are from the MIMIC-II dataset.

mitigated against this by only providing low RR estimates if both time- and frequency-domain RR estimates agreed. Fusion was found to be beneficial below 18 bpm for the ECG, and 16 bpm for the PPG. In this algorithm agreement was assessed between RR estimates to determine whether to output a final RR estimate. More advanced techniques consider the quality of respiratory signals from which RR estimates were derived to determine whether to output a final RR, and may warrant further investigation [154, 155, 199].

Further testing and refinement of the algorithm is required prior to implementation in clinical practice. Firstly, the algorithm should be assessed using data which is representative of acutely-ill hospitalised patients. The dataset most representative of acutely-ill patients in this study was the MIMIC-II dataset. However, even this dataset may not be fully representative

of acutely-ill patients, since some of the critically-ill patients in this dataset were receiving mechanical ventilation, which may affect the performance of RR algorithms [341, 346]. They also received drugs such as inotropes, beta-blockers and vasoactive agents which may impact the physiological mechanisms which cause respiratory modulation of the ECG and PPG (see Section 4.2). Secondly, the optimal design specification of an RR algorithm for continuous monitoring of hospitalised patients is not yet clear. The relative importance of precise RR estimates, avoiding highly erroneous estimates (both captured in C), and providing RR estimates for a high proportion of windows (*prop*) is not yet known. Thirdly, there was variation in the algorithm's performance between different subjects, and some variation with the reference RR. Further investigation is required to determine whether these variations in performance affect the clinical utility of the algorithm.

This investigation was conducted using the **RRest Toolbox** of algorithms (described in Section 4.6.3). The toolbox is publicly available (see the accompanying tutorial for an introduction [6]), and has several features which may be of benefit to other researchers. The toolbox can be used to perform equivalent analyses across multiple datasets (four were used in this investigation). To do so, scripts are provided to download and appropriately format each dataset for use with the toolbox. In addition, several methods are provided to precisely derive reference RRs from each dataset, whether using a reference respiratory signal or breath annotations. Three different methods were used to derive reference RRs in this investigation. Furthermore, the toolbox provides a standardised framework for assessing the performance of RR algorithms. This allows the performances of algorithms reported in this investigation to be directly compared to those of algorithms used in other investigations conducted with the toolbox, such as the assessment of previously proposed RR algorithms in Chapter 6.

This study highlighted the importance of selecting an appropriate dataset with which to assess the performance of RR algorithms. Two datasets containing data from hospitalised patients were used: CapnoBase and MIMIC-II. However, the conclusions were drawn primarily from analyses conducted on the MIMIC-II dataset. This dataset was prioritised since it is more representative of acutely-ill adult patients. In addition, the algorithms did not perform as well on this dataset, making it a worst-case scenario. It is important that datasets used for assessment of RR algorithms are as representative as possible of the clinical setting in which the algorithms will be used. Further work is now ongoing to assess the performance of the novel algorithm on the RRest-acute dataset, which is designed to be as representative as possible of acutely-ill hospitalised patients (see Section 4.6.1).

7.6 Final Remarks

In this chapter a novel RR algorithm for continuous monitoring of acutely-ill hospitalised patients was designed and optimised. The algorithm uses a frequency-domain technique to estimate higher RRs, and both time- and frequency-domain techniques to estimate lower RRs. In doing so it utilises the high precision of many estimates produced using frequency-domain techniques, whilst mitigating against the underestimates often produced by these techniques. The algorithm's performance was compared to that of the best-performing algorithm from the literature, identified in Chapter 6. The algorithm was found to provide greater precision than the control algorithm, at the expense of providing RR estimates for a smaller proportion of windows of input signal. It was found to perform sufficiently well for continuous monitoring when used with the ECG, although not when used with the PPG.

The results from this study suggested that the novel algorithm's performance may vary between patients, and across different RRs. Therefore, further work is required to determine whether the novel algorithm performs as expected when used with wearable sensors in the hospital setting.

Chapter 8

Detecting Deteriorations using Wearable Sensors

This chapter presents a study investigating the potential utility of wearable sensors for detecting deteriorations using the LISTEN Processed Dataset (Chapter 3). The respiratory rate (RR) algorithm presented in Chapter 7 was used to estimate RR continuously from the electrocardiogram (ECG) signals monitored by the wearable sensors. This provided three continuously monitored parameters: heart rate (HR), arterial blood oxygen saturation (SpO_2), and RR. The parameters were median-filtered to eliminate short-term variability. Three algorithms were then used to assess the likelihood of deterioration at each time point, and a windowing scheme was used to activate and deactivate alerts. The results were used to assess the performance of algorithms to detect deteriorations when applied to wearable sensor data compared to routine, intermittent physiological measurements.

8.1 Introduction

It has been widely suggested that wearable sensors could be used to provide early warning of clinical deteriorations [61, 66, 69, 91, 347, 348]. Recent developments in wearable sensor technology have made them particularly suited to this application (as detailed in Section 2.3). Developments have included: broadening the range of physiological parameters which wearable sensors can monitor [29]; reducing their size and weight [66, 67]; increasing their battery life [68]; and automating rejection of low quality measurements [67, 124]. However, relatively little

attention has been paid to the design of algorithms to detect deteriorations from wearable sensor data, prompting a clinical response [348]. This is a key step, since no monitoring device can improve patient outcomes unless coupled with an effective clinical intervention [349]. There is no commonly accepted approach for the design of algorithms to detect deteriorations from wearable sensor data, and there is little evidence for the clinical utility of such algorithms. Therefore, further research is required to optimise algorithm designs, and to determine whether they provide earlier warning of deteriorations than current practice [348].

A significant barrier to research in this area is the difficulty of acquiring wearable sensor data from sufficiently large numbers of patients. Clinical trials to acquire datasets using wearable sensors have suffered from poor data coverage rates [75] and short monitoring durations (see Section 2.4). In contrast, the LISTEN Processed Dataset, described in Chapter 3, is a uniquely comprehensive wearable sensor dataset. It consists of wearable sensor data, routine physiological measurements, and labels of adverse events acquired from 226 acutely-ill patients (see Section 3.4). The wearable sensors monitored HR and SpO_2, as well as acquiring ECG and PPG signals continuously (as described in Table 3.1). Consequently, when an algorithm is used to retrospectively estimate RR from the ECG or PPG signals, three physiological parameters can be continuously derived from the wearable sensor data: HR, SpO_2 and RR. This provides a comparable set of parameters to those acquired by existing wearable sensors for use with acutely-ill patients. For instance, the SensiumVitals ® patch acquires HR, RR and temperature [68]; and the Philips Biosensor acquires HR, RR, temperature, and activity measures [350]. Although the LISTEN Processed Dataset is subject to some limitations, as described in Section 3.5, it is a valuable resource for developing and evaluating algorithms to detect deteriorations from wearable sensor data.

This chapter presents a study in which the LISTEN Processed Dataset was used to assess the potential utility of wearable sensors for early detection of deteriorations. The aim of this pilot study was to assess the performance of algorithms to detect deteriorations when applied to wearable sensor data compared to routine, intermittent physiological measurements. The algorithms used to detect deteriorations are presented in Section 8.2. Section 8.3 describes the methods used to evaluate the algorithms using the LISTEN Processed Dataset. The results are described in Section 8.4. The implications of the results, and the potential utility of the algorithms for use in wearable sensors are discussed in Section 8.5.

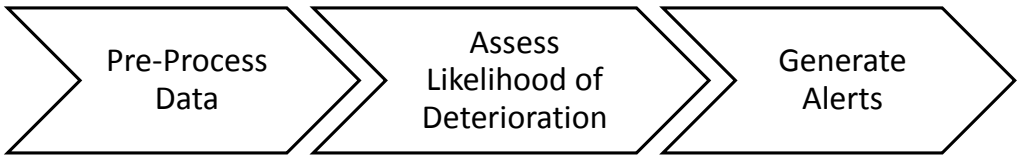

FIGURE 8.1: The stages of an algorithm to detect deteriorations from wearable sensor data.

8.2 Algorithms to Detect Deteriorations

Algorithms to detect deteriorations from wearable sensor data generally consist of three stages, as illustrated in Figure 8.1. The first stage, pre-processing of wearable sensor data, involves cleaning the data corresponding to each physiological parameter. This is important since the data are particularly susceptible to noise. Reasons for this include [4]: (i) movement artefact is common since wearable sensors are used with ambulatory patients; (ii) wearable sensor data are acquired continuously, regardless of whether a patient is active, causing physiological parameters to vary from their baseline values at rest; (iii) sensors can become partially detached resulting in poor contact. Therefore, data pre-processing is important to ensure that the data are representative of a patient's underlying physiological state. The second stage, assessing the likelihood of deterioration, usually involves using a novelty detection technique to compare a set of wearable sensor data to a model of normal physiology (physiology in the absence of adverse clinical events). The third stage, generating alerts of deteriorations, consists of comparing time series of likelihoods of deterioration to a set of criteria to determine whether a new alert should be raised, or in the case of an ongoing alert, whether it should be stopped. Each stage of the algorithm is important in ensuring that alerts of deteriorations are as useful as possible.

The algorithms used to detect deteriorations in this study are now described.

8.2.1 Data Pre-Processing

The first pre-processing step was to filter the HR, SpO_2 and RR parameters to eliminate short-term variability, and produce more stable values from which the likelihood of deterioration could be assessed. Some short-term variability was evident in the LISTEN Processed Dataset, despite the prior use of signal quality indices to eliminate any parameters derived from low quality signals (see Section 3.3). An example of this is shown in Figure 8.2 (where crosses indicate the raw 1 Hz measurements). Median filtering appears to be a reasonable technique for eliminating short-term variability from wearable sensor data since it has been widely used to attenuate

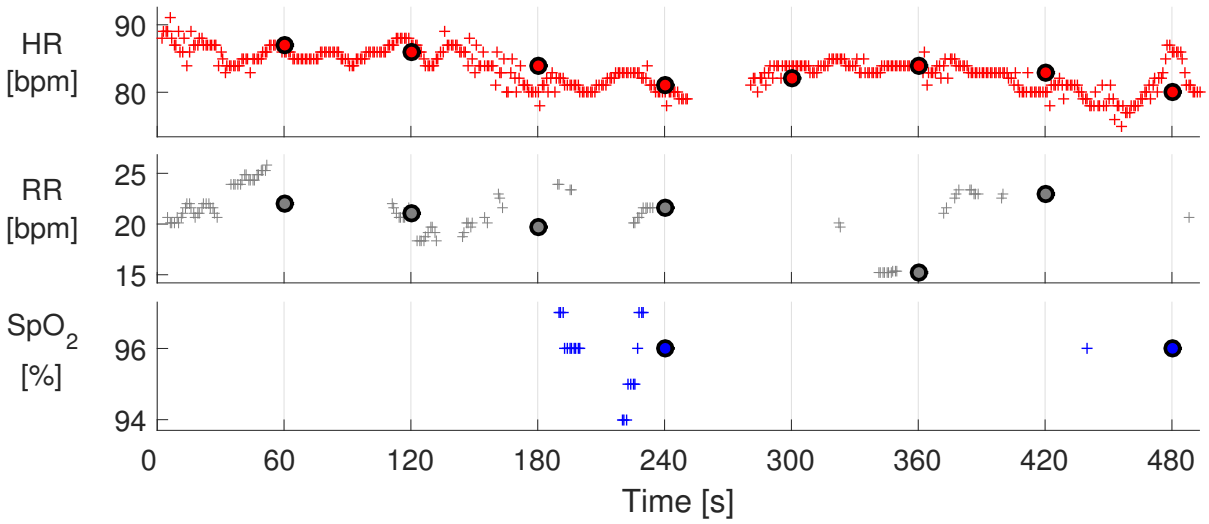

FIGURE 8.2: Pre-processing of wearable sensor data using a median filter: the raw HR, RR and SpO_2 measurements at 1 Hz were median-filtered to produce filtered values once per minute. Raw values at 1 Hz are shown by crosses, and median-filtered values are shown by large dots. In this example HR values were available during each minute of monitoring, and therefore a median-filtered value was calculated for each minute. Median-filtered RR values were calculated for all minutes except those finishing at 300 and 480 s, in which there were no raw values available. Raw SpO_2 values were only available for two windows, resulting in only two median-filtered values.

high frequency noise from physiological time series [351–353], including as a pre-processing step prior to detection of deteriorations [354]. Median filters have the advantage of retaining edges in a signal which are indicative of a sustained change in physiological state [351]. Therefore, the parameters (originally sampled at 1 Hz) were median-filtered to provide filtered values once per minute (using a filter window duration of one minute), as shown in Figure 8.2 (large dots indicate filtered values). The benefit of median-filtering is demonstrated by the raw SpO_2 values between 180 and 240 s in this figure, which ranged from 94 - 97 % over the course of only one minute. The use of median-filtering across this window eliminated the short-term variability. Without this step the high level of variability in the raw values could result in variability in the assessed risk of deterioration over a short period of time, potentially causing false alerts.

The second pre-processing step was to generate a complete set of data at each time point with which to assess the likelihood of deterioration. The five routinely measured vital signs (HR, SpO_2, RR, systolic blood pressure - SBP, and temperature - temp) were used, as well as a binary variable indicating whether or not supplemental oxygen was being administered (O_2). There were gaps in the HR, SpO_2 and RR data at times when these parameters were either not acquired, or were of low quality. Furthermore, SBP, temp and O_2 were only acquired during intermittent, manual measurements as they were not monitored by the wearable sensors. A

sample-and-hold algorithm was used to impute the most recent measurement at time points when a parameter was not measured. This provided a complete set of parameters, from which the likelihood of deterioration could be assessed at all time points.

8.2.2 Assessing the Likelihood of Deterioration

Novelty detection algorithms have previously been used to assess the likelihood of deterioration from wearable sensor data [348]. They are particularly suitable for detection of deteriorations since they can be trained using *normal* physiological data, which is easier to acquire since most patients do not suffer from major adverse events. In contrast, it is relatively difficult to acquire the *abnormal* physiological data which would be required for two-class classification [355]. Novelty detection algorithms used with wearable sensors take one of three broad approaches: (i) comparison of a single-parameter to a threshold indicating abnormality [85]; (ii) continuously updating a traditional EWS using wearable sensor data [114]; and, (iii) a more complex novelty detection algorithm requiring electronic calculation, such as kernel density estimation (KDE), support vector machines, Gaussian processes and Gaussian mixture models [348]. The first approach, that of using a single parameter in isolation, is not considered in this work since deteriorations are expected to be manifested across multiple parameters [44]. Commercially available implementations of the latter two approaches have been described in Section 2.5.2. The implementations of these approaches used in this study are now described.

8.2.2.1 Abbreviated Early Warning Scoring

The first algorithm used to assess the likelihood of deterioration was a continuous implementation of the National Early Warning Score (NEWS, described in Section 2.2). The NEWS is used in the UK to detect deteriorations from intermittent physiological measurements [28]. It is calculated from seven intermittently measured vital signs (see Table 2.1 for further details): HR, RR, SpO_2, systolic blood pressure (SBP), temperature (temp), oxygen therapy (O_2) and level of consciousness (LOC). In this study $NEWS_a$, an abbreviated version of the NEWS, was used to assess the likelihood of deterioration. The NEWS was abbreviated by removing the need for LOC, since this parameter was not acquired in the Listen Trial. This means that $NEWS_a$ is an underestimate since any potential contribution to the likelihood of deterioration from LOC is omitted.

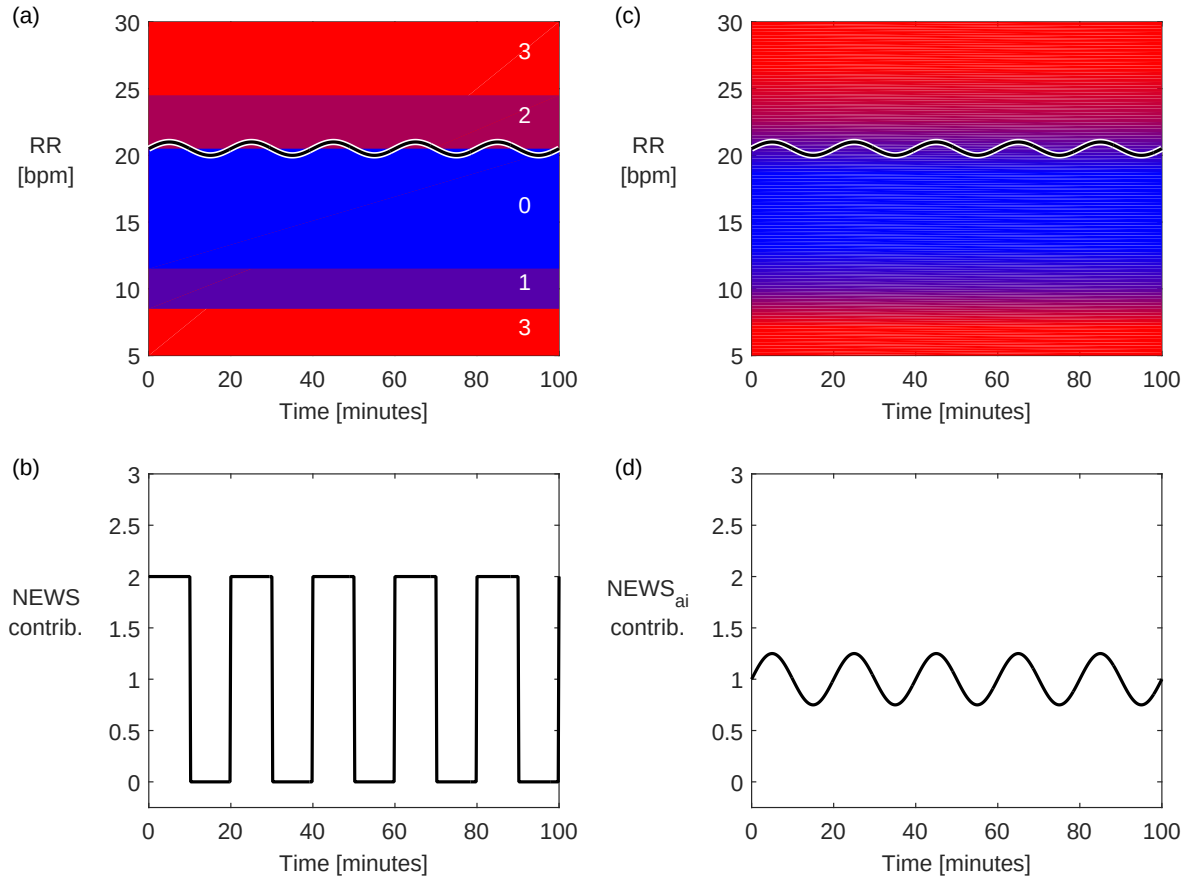

FIGURE 8.3: The benefit of NEWS$_{ai}$, an interpolated National Early Warning Score (NEWS): Wearable sensor data is subject to variability due to its high sampling frequency. The NEWS is a step function of the input parameters, as shown in the background of (a). Therefore, a small fluctuation in RR, perhaps of only 1 bpm as shown in (a), causes an unreasonably large fluctuation in the NEWS as shown in (b). This may cause false alerts. In contrast, when a continuous function such as NEWS$_{\text{ai}}$ is used as shown in (c), small changes in RR only result in small changes in the contribution to NEWS$_{\text{ai}}$. This may reduce the frequency of false alerts compared to a step function.

8.2.2.2 Interpolated Early Warning Scoring

The second algorithm, NEWS$_{\text{ai}}$, was a further modification of the NEWS, designed specifically for use with wearable sensor data. The NEWS$_{\text{a}}$ algorithm was interpolated to provide a continuous output rather than a quantised output. The NEWS algorithm assigns an integer score to each vital sign, and the overall score is obtained by summing the integer scores for each vital sign. For instance, a score of 0 is assigned to the RR if $12 \leq \text{RR} \leq 20$ bpm, and a score of 2 if $21 \leq \text{RR} \leq 24$ bpm. However, if the vital sign value happens to lie close to a scoring threshold then natural variability in physiology will cause unreasonably large fluctuations in the NEWS. For instance, if a patient's RR is approximately 20 bpm, then every time the RR rises to 21 bpm the NEWS will increase by 2, as demonstrated in Figure 8.3 (a) and (b). This gives the mis-

leading impression of a marked increase in the likelihood of deterioration for a relatively small physiological change. Each vital sign's contribution to NEWS_{ai} was calculated using a linear interpolation of the algorithm's current integer contributions to avoid these large fluctuations. The application of this approach to calculating the contribution of RR to the overall score can be seen in the background of Figure 8.3 (c) and (d).

8.2.2.3 Kernel Density Estimation

The third algorithm, KDE, calculated the level of abnormality (termed novelty) of a set of physiological parameters (a pattern) through comparison with a model of normal physiology. To do so a training dataset of normal physiology was extracted from the LISTEN dataset. The likelihood of deterioration was assessed in four steps: (i) the training dataset was normalised so that each parameter had zero mean and unit standard deviation; (ii) the dataset was reduced to a smaller set of representative features for ease of computation; (iii) a model of normality was constructed from the smaller set of features using KDE; (iv) the model of normality was used to evaluate the novelty of a set of physiological parameters. The methodology used for each step is now explained. Further information on the approach can be found in [354, 356].

The first step was to normalise each parameter in the training dataset to have zero mean and unit standard deviation [354]:

$$x_n = \frac{x - \mu_g}{\sigma_g} \quad , \tag{8.1}$$

where x_n is the normalised value of measurement x, and μ_g and σ_g are the mean and standard deviation of the parameter in the training data. The parameters were normalised in preparation for the use of KDE later in the procedure, in which a single variance was assumed to describe the deviation of all the input parameters from their means. Extremely abnormal patterns were identified as those which contained a value $> 5\sigma_g$. These patterns were excluded from the training dataset to avoid including highly imprecise measurements.

The second step was to reduce the normalised training dataset to a smaller set of representative patterns. This was necessary because the computational cost of comparing an incoming pattern to all the individual patterns in the training dataset would be overly high. The patterns in the dataset were grouped into $k = 500$ clusters as suggested in [354, 356, 357]. This was performed using the k-means^{++} algorithm, which identifies k central patterns, attempting to minimise the squared distance between each original pattern and its closest centre [358]. Outlying clusters

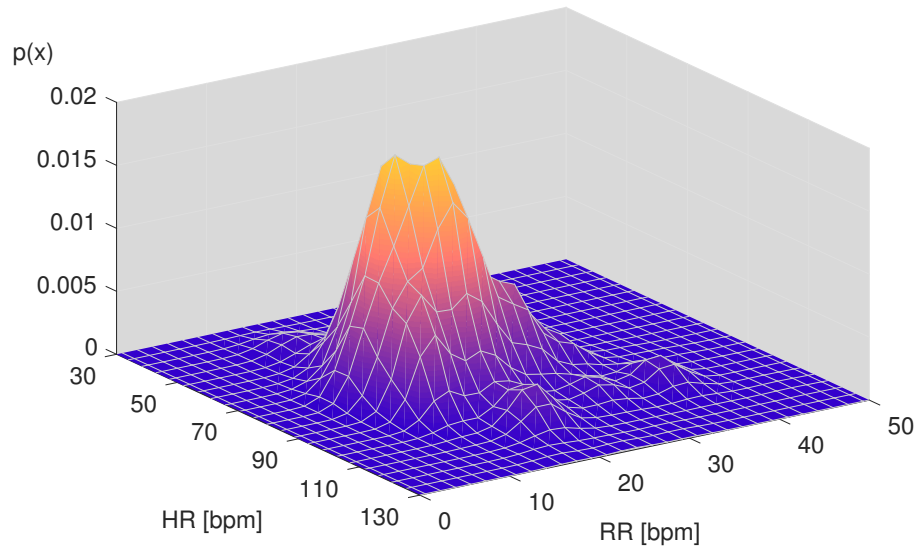

FIGURE 8.4: Using kernel density estimation (KDE) to estimate the probability density function (PDF) of a multivariate dataset: An estimated PDF, $p(\mathbf{x})$, is shown, where the two-dimensional feature vector $\mathbf{x}$ consists of HR and RR. The five-dimensional model of normal physiology estimated using KDE in this study also contained SpO_2, SBP and temp.

were excluded to retain only the *most normal* clusters as suggested in [354, 356, 357]. The exclusion of outlying clusters is intended to ensure that transient departures from normality exhibited by *normal* patients are not included in the model. A total of 15 % of clusters were excluded in line with [354], leaving 425 clusters. Two methods have been used in this field for identifying outlying clusters for exclusion: clusters can be ranked according to their distance from the mean, and those furthest from the mean excluded [354]; or, clusters can be ranked according to their weights (numbers of patterns allocated to each cluster), and those with the lowest weights excluded. The weighted method was used in this study because it was found to provide superior performance in [359].

The third step was to construct a model of normality using KDE. The Parzen window kernel density method was used [360] to estimate the probability density function (PDF), $p(\mathbf{x})$, of the training dataset, where $\mathbf{x}$ is a pattern at a single time point. An example is shown in Figure 8.4. The Parzen window method consists of locating a Gaussian kernel at each of the cluster centres. $p(\mathbf{x})$ is then estimated by evaluating the sum of the contributions of each kernel to the overall probability distribution [357]. This results in the following equation for $p(\mathbf{x})$ (derived in [354]):

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\sigma^D} \sum_{i=1}^{N} w_i \exp\left(\frac{-\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}\right) \quad , \tag{8.2}$$

where N is the number of clusters used to represent the training dataset, D is the number of

dimensions in the dataset, w_i is the cluster weight ($w_i = n_i / \sum_{i=1}^{N} n_i$, where n_i is the number of patterns associated with a cluster) and σ is the bandwidth (standard deviation) of the isotropic kernels. It is important to select an appropriate value for σ, which acts as a smoothing function, since too small a value will result in overfitting the training data, and too large a value will not capture local variation [357]. Several methods have been proposed to select an appropriate value of σ [357, 359]. In this work the optimal value of σ was selected by maximising the log-likelihood of the training patterns (for $0.02 \leq \sigma \leq 1.00$). This method for optimising σ was used as it has previously been shown to perform well in a comparison of different methods across several datasets [359].

The final step was to use the model of normality to evaluate the novelty (level of abnormality) of a pattern. Novelty has previously been calculated using [70, 354, 357]

$$\text{novelty}(\mathbf{x}) = \log_e \frac{1}{p(\mathbf{x})} = -\log_e p(\mathbf{x}) \quad , \tag{8.3}$$

which increases as $p(\mathbf{x})$ decreases. An increase of 1 in the novelty compared to the previous value indicates that the current $\mathrm{p}(\mathbf{x})$ is e times smaller than the previous value. The novelty score is offset so that the maximum of $\mathrm{p}(\mathbf{x})$ results in a novelty of approximately zero. This was performed by calculating the novelty at $\mathbf{x} = [0, 0, 0, 0, 0]$, as suggested in [354]. This value was then subtracted from all novelty calculations.

8.2.3 Generating Alerts

Techniques for generating alerts from estimated likelihoods of deterioration have received relatively little attention in the literature. The simplest method is to generate an alert every time the likelihood breaches a pre-defined threshold. However, this method can result in excessive activation and deactivation of alerts due to natural variation in physiology. A more advanced method was proposed in [354] for use with static patient monitors to activate and deactivate alerts in an attempt to reduce the frequency of false alerts. This consisted of activating an alert only if the likelihood was above a threshold value for at least 80 % of the previous five minutes' data [115]. Similarly, an alert was only deactivated if the likelihood was below the threshold for at least 80 % of the previous five minutes' data. This method has since been used in a clinical study [361]. The benefits of this more advanced method are illustrated in Figure 8.5. This shows the use of each of the methods to generate alerts for a patient's entire

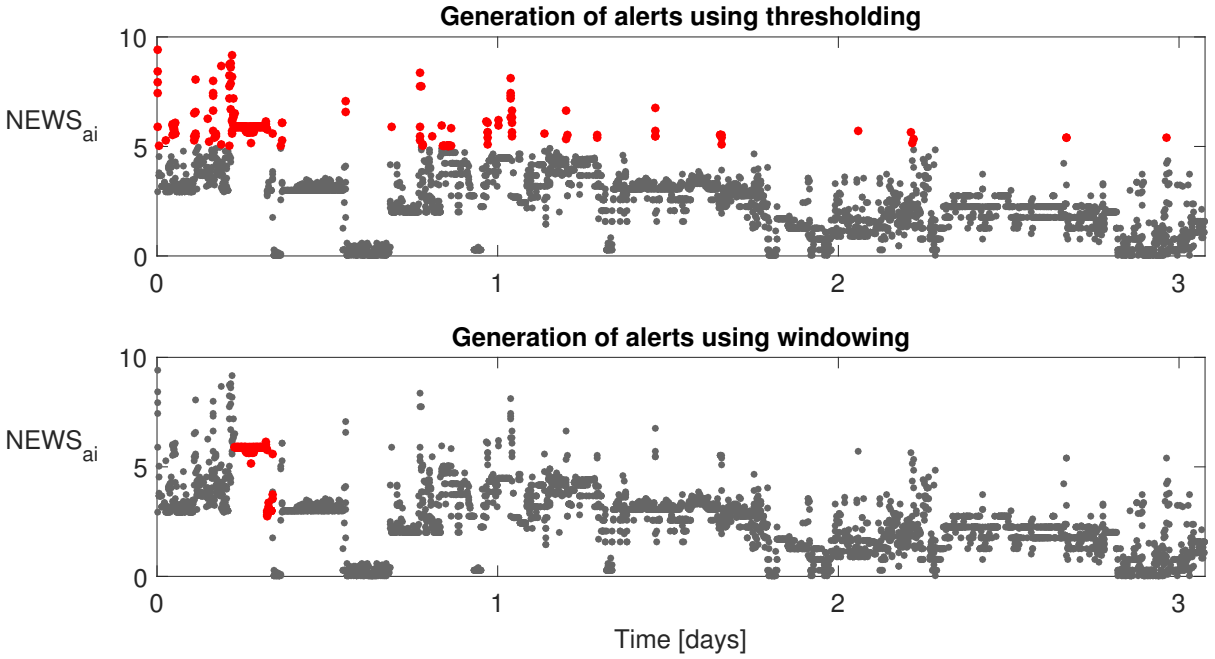

FIGURE 8.5: A comparison of methods for generating alerts of deteriorations from wearable sensor data: The results of two different methods for generating alerts are shown, where red dots indicate an alert. In the upper plot an alert was activated when the likelihood of deterioration, NEWS_{ai}, was greater than a threshold value (5 in this case). In the lower plot an alert was activated if at least 80 % of the NEWS_{ai} values in the last 30 minutes were greater than 5, and only deactivated when at least 80 % of the NEWS_{ai} values in the last 30 minutes were less than 5.

duration of wearable sensor monitoring (3.1 days). The simple method raised alerts throughout the duration of monitoring, until discharge from hospital, despite the patient not suffering from any adverse events. In contrast, the more advanced method only generated alerts in the first half day of monitoring, when physiology is more likely to be unstable [51]. Thereafter the false alerts caused by transient changes in physiology were suppressed by the advanced method.

8.3 Methods

8.3.1 Respiratory Rate Estimation

RRs were estimated retrospectively from the ECG signals in the LISTEN Processed Dataset using the algorithm proposed and assessed in Chapter 7 (at 1 Hz using 32 s windows). Figure 8.6 shows an example of RRs estimated from the ECG, alongside those acquired simultaneously as part of routine monitoring using impedance pneumography (ImP) in critical care. The performance of the RR algorithm on this particular dataset is currently being assessed by clinical collaborators using the RRest-acute, RRest-afib and RRest-vent datasets described in Section

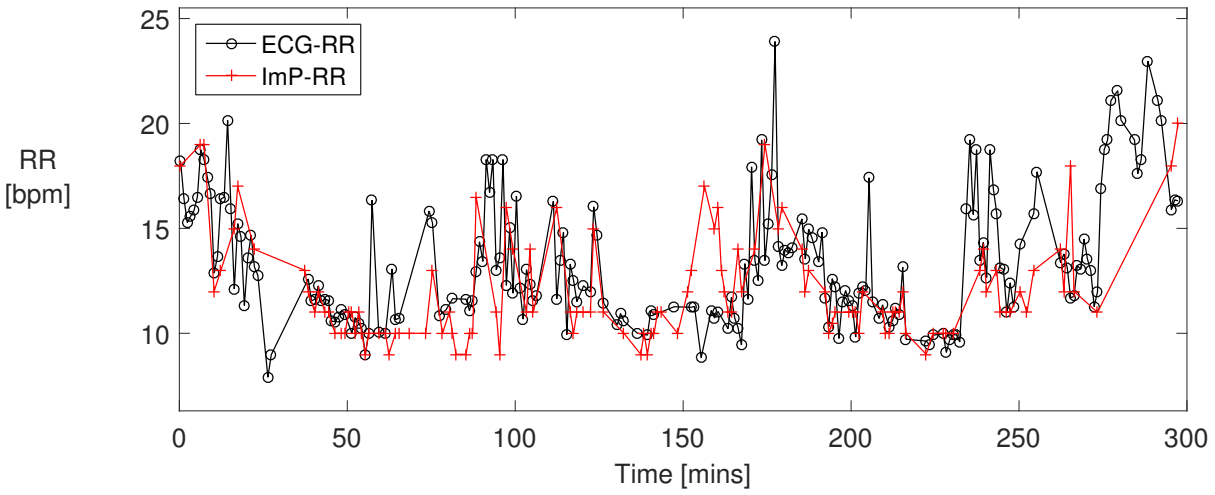

FIGURE 8.6: RRs estimated retrospectively from the ECG signals in the LISTEN Processed Dataset (shown in black), alongside those acquired during routine critical care monitoring using impedance pneumography (ImP, shown in red). The RRs shown have been median-filtered to produce (at most) one estimate per minute.

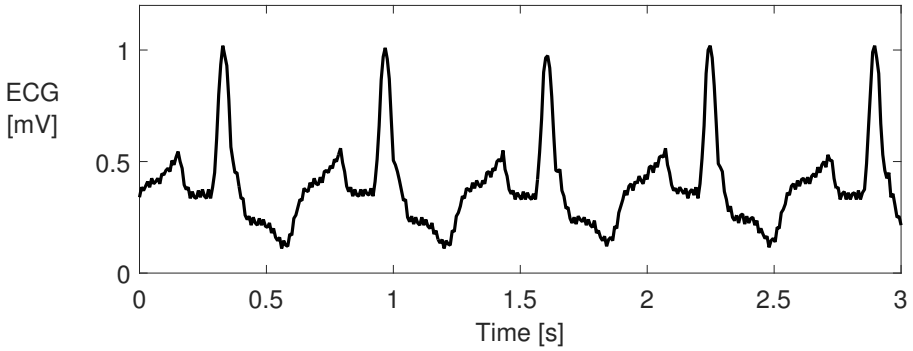

FIGURE 8.7: An ECG signal on which the RR algorithm did not perform well. In this case, the suboptimal performance may have been caused by abnormal ECG morphology, resulting in a reduction of the quality of respiratory signals extracted from the ECG using feature measurement techniques.

4.6.1. Preliminary results showed high performance. However, the algorithm was not flawless, with poorer performance seen in unusual situations such as with the ECG signal shown in Figure 8.7, which exhibits abnormal morphology. The abnormal morphology may have resulted in reduced quality of respiratory signals extracted from the ECG using feature measurement techniques, leading to imprecise RR estimates.

8.3.2 Patient Classification

The LISTEN Processed Dataset contains data from 226 acutely-ill hospitalised patients recovering from cardiac surgery (see Section 3.4 for further details). Only the 184 patients who had wearable sensor data available after median-filtering were included in this study. Patients were

TABLE 8.1: Adverse events used to categorise LISTEN patients

Classification	Events
Severe Adverse Event (SAE)	readmission to critical care, death, cardiac arrest, unplanned return to surgery, 'do not resuscitate' order
Adverse Event (AE)	stroke or transient ischaemic attack, seizure, reintubation, pneumothorax, ischaemic ECG changes, insertion of pleural drain or aspiration, insertion of other drain, commencement of inotropes or intra-aortic balloon pump, deep wound infection, cardiac failure, bowel obstruction, bleeding, sepsis, acute myocardial infarction, altered mental state, acute kidney injury, atrial fibrillation requiring treatment, non-invasive ventilation or continuous positive airways pressure ventilation, unplanned cardiac pacing or DC cardioversion, ventricular tachycardia, torsade de pointes, transient loss of consciousness

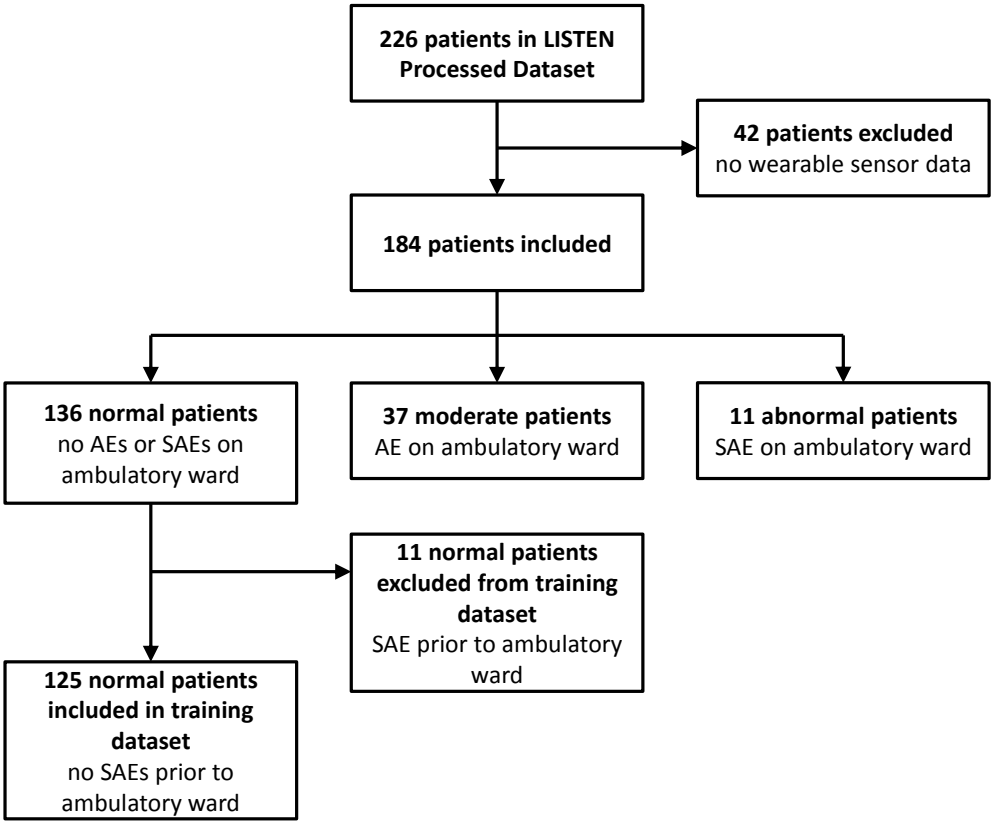

FIGURE 8.8: Flow of patients through the study

divided into three categories: (i) normal (did not experience an adverse event - AE - on the ambulatory ward); (ii) moderate (experienced an AE on the ambulatory ward); and, (iii) abnormal (experienced a severe adverse event - SAE - on the ambulatory ward). Table 8.1 lists the events which were considered to be AEs and SAEs. Figure 8.8 shows the numbers of patients assigned to each category.

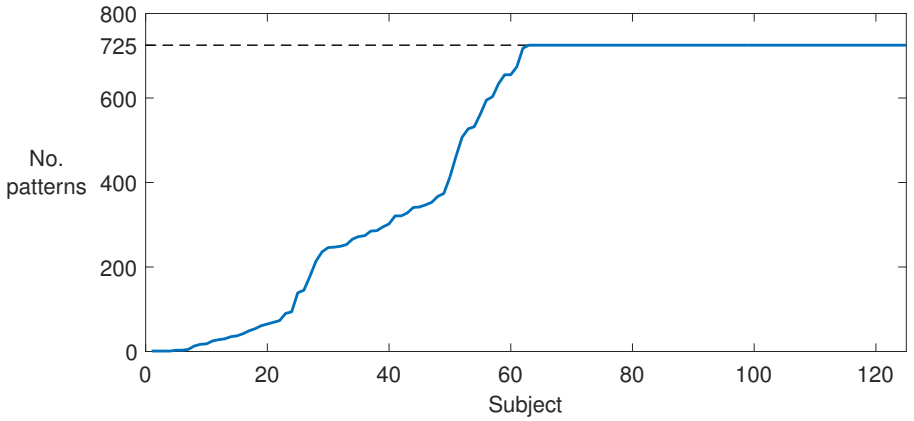

FIGURE 8.9: The number of patterns contributed by each patient to the training dataset: the number of patterns per patient was limited to 725, the median number available per patient. Therefore, 62 patients contributed < 725 patterns, whilst the remainder each contributed 725 patterns.

8.3.3 Extracting a Training Dataset

A training dataset of normal physiology was extracted from the LISTEN Processed Dataset. Only normal patients were considered for inclusion in the training dataset. Those which experienced a SAE prior to admission to the ambulatory ward were excluded since any physiological derangement caused by the SAE could have remained during wearable sensor monitoring. Consequently, 125 normal patients were included in the training dataset (see Figure 8.8).

Many patterns were available from each patient since wearable sensor data were typically acquired for several days, sampled once per minute after median-filtering. Patterns were only included in the analysis if measurements of HR, SpO_2 and RR had been acquired by a wearable sensor during the minute from which a pattern was generated. Furthermore, the number of patterns contributed by any individual patient was limited to the median number of patterns available from each patient. This resulted in a median and maximum of 725 patterns per patient (see Figure 8.9). A subset of patterns were identified for each patient for whom more than this number of patterns were available. This subset consisted of 725 patterns equally spaced from the first to the last pattern of each patient's recording. The resulting training dataset, containing 6.1×10^4 patterns, was balanced by ensuring that no individual patient contributed an excessive number of patterns, and by ensuring that patterns were taken from throughout the time for which wearable sensor data were acquired.

The distributions of individual parameters in the training dataset are shown in Figure 8.10. The histograms of temp and RR show approximately normal distributions. SBP and HR exhibit

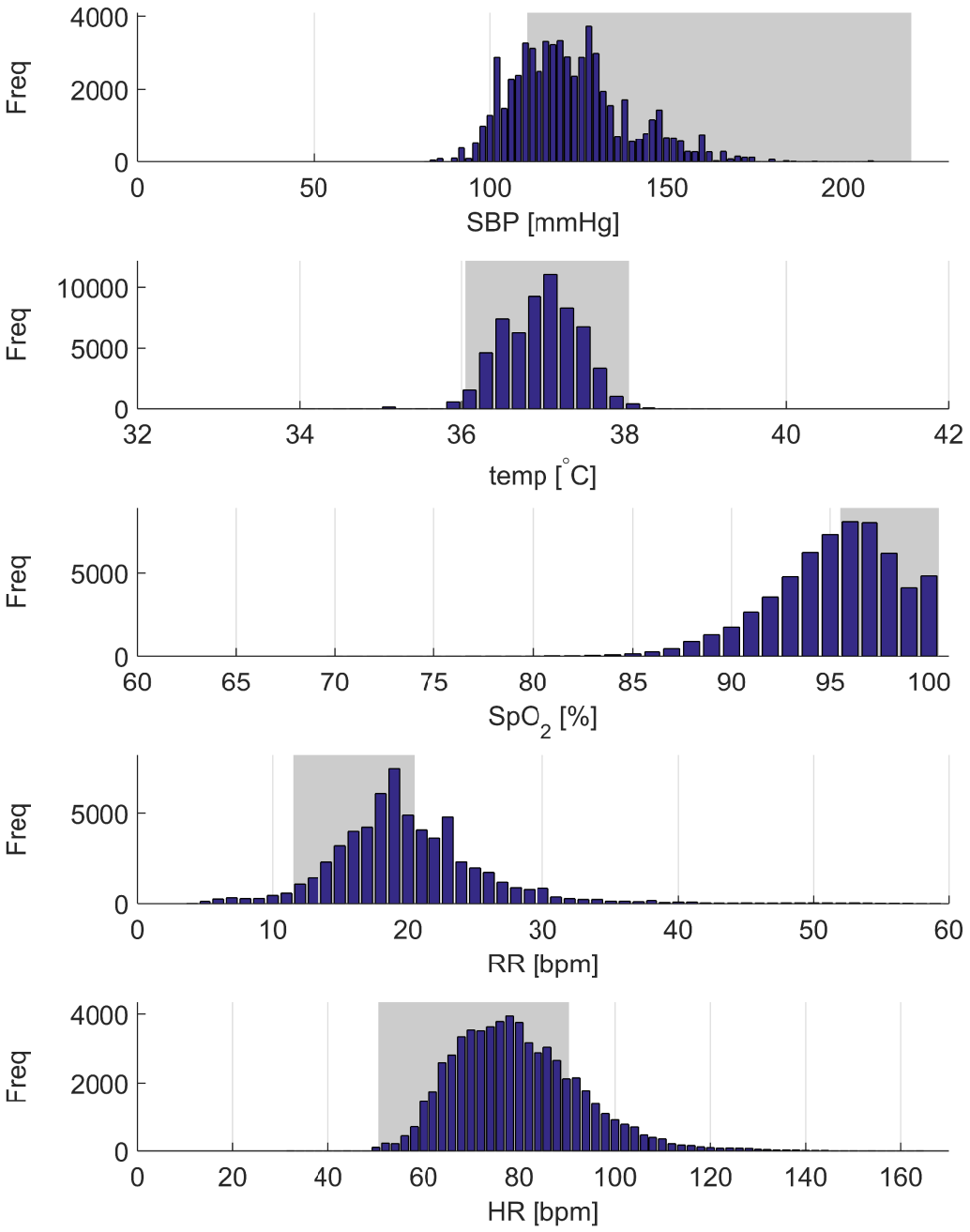

FIGURE 8.10: Histograms of individual parameters in the training dataset. Ranges shaded grey show values deemed to be normal according to the NEWS.

positive skews, and SpO_2 exhibits a large negative skew (as expected since SpO_2 cannot exceed 100 %). Normal ranges of values at rest, shaded in grey, are those which score 0 in the NEWS, indicating minimal likelihood of deterioration (see Table 2.1 for details). The histogram for temp coincides well with the normal range. In contrast, a substantial proportion of the measurements of the remaining parameters lie outside of their normal ranges. SBP and SpO_2 tended to be lower than normal, and RR and HR were slightly higher than normal. This may have been due to: (i) the post-operative population, whose physiology may have been deranged by the major

TABLE 8.2: Normalisation of physiological parameters.

Parameter	mean, μ_g	standard deviation, σ_g
HR [bpm]	79.1	13.4
RR [bpm]	20.0	5.7
SpO_2 [%]	95.2	3.2
SBP [mmHg]	122.1	16.2
temp [°C]	36.9	0.5

surgery only days before these measurements; and, (ii) HR, RR and SpO_2 measurements being acquired by wearable sensors in all conditions, rather than only at rest.

8.3.4 Algorithms to Detect Deteriorations

Three algorithms were implemented to detect deteriorations as described in Section 8.2. The data steps for pre-processing and generating alerts were common to all three algorithms. The algorithms differed in their methodology for assessing the likelihood of deterioration, using the $NEWS_a$, $NEWS_{ai}$ and KDE approaches respectively. Details of the methods which are specific to this study are now described.

The training dataset was normalised to create a model of normal physiology for the KDE approach. The mean and standard deviation values calculated for each parameter are provided in Table 8.2. Extremely abnormal patterns, accounting for 0.54 % of the training patterns, were excluded to avoid including highly imprecise measurements. The majority of these patterns (72.0 %) were abnormal because of extremely high RR measurements ($>$ 48.5 bpm), which are likely to be imprecise. The remaining patterns were excluded due to extremely low SpO_2 measurements (18.9 %), extremely high SBP measurements (6.7 %), and extremely high HR measurements (2.4 %). Following clustering of the data, the optimal bandwidth of the kernels for use in the KDE was found to be $\sigma = 0.34$. The KDE novelty value at $\mathbf{x} = [0, 0, 0, 0, 0]$ was 4.77, which was subtracted from the novelty calculations when using the KDE approach.

The generalised methodology for generating alerts described in Section 8.2.3 was adapted for use with wearable sensors in this study. An alert was activated (or deactivated) if a novelty score had been calculated for at least 50 % of the previous 30 minutes' data, and at least 80 % of the scores were greater than (or less than) the relevant threshold (*e.g.* $k_{w,NEWSai}$ for the $NEWS_{ai}$ algorithm, where subscript w indicates a threshold optimised for wearable sensor

data). Optimal thresholds were identified as those which limited the frequency of false alerts to a maximum of one per day, as suggested in [11].

8.3.5 Comparison to Routine Practice

Patients in the LISTEN dataset had routine, manual physiological assessments approximately 4 times per day (as reported in Table 3.6). Three early warning scores were calculated from these intermittent data as benchmarks against which to compare the algorithms applied to wearable sensor data. All three approaches used to assess the likelihood of deterioration with wearable sensor data were also used with the intermittent data. When using intermittent data an alert was raised immediately if the score breached an alerting threshold (*e.g.* $k_{\text{i,NEWSai}}$ for the NEWS_{ai} algorithm, where subscript i indicates a threshold optimised for intermittent data). Optimal thresholds were set in the manner as for the wearable sensor data.

8.3.6 Performance evaluation

To evaluate the performances of the six methods (three algorithms applied to wearable sensor data, and three applied to intermittent data), data from each subject were segmented into windows of one hour duration. Reference classifications for each window were assigned as follows. Windows were classified as negative if they were not proceeded by an AE or a SAE. Windows were classified as positive-AE, or positive-SAE, if the first AE, or the first SAE, following admission to the ambulatory ward occurred within 48 hours of the start of the window. The novelty scores in each window were classified as positive if the window contained an alert.

Performances were initially assessed using the area under the receiver-operator curve (AUROC) statistic, since this has been widely used to assess the performance of algorithms to detect deteriorations [11, 23, 45, 53, 130, 362]. The algorithm with the highest AUROC was selected for further analysis. The performance of the selected algorithm was then assessed when using the optimal threshold to generate alerts. The widely used sensitivity and specificity metrics were reported to allow comparison with other algorithms from the literature. However, it has recently been noted that the specificity is not highly informative in this setting, since physiological deterioration has an extremely low prevalence [363]. Therefore, the positive predictive value (PPV), and proportion of scores which alerted were also reported.

TABLE 8.3: Data characteristics: The number of 1-hour windows classified as negative (those which were not proceeded by an adverse event - AE or severe adverse event - SAE) and positive (those which were proceeded by an AE or SAE within 48 hours).

	Window classification, n (%)	
Events	Negative	Positive
SAEs	17,872 (98.5)	263 (1.5)
AEs	17,872 (95.0)	935 (5.0)

TABLE 8.4: Performances of algorithms to detect deteriorations when applied to wearable sensor data and to intermittent data, evaluated using the area under the receiver-operator curve (AUROC).

		AUROC	
Events	Algorithm	Wearable sensor data	Intermittent data
SAEs	$NEWS_a$	0.756	0.753
	$NEWS_{ai}$	0.780	0.794
	KDE	0.731	0.677
AEs	$NEWS_a$	0.706	0.657
	$NEWS_{ai}$	0.727	0.686
	KDE	0.684	0.606

8.4 Results

8.4.1 Dataset Characteristics

Table 8.3 shows the numbers of windows classified as negative and positive when evaluating the performance of algorithms for detection of SAEs and AEs. There were low incidences of positive windows, indicating windows within the 48 hours prior to an adverse event, when detecting both SAEs (1.5 %) and AEs (5.0 %).

8.4.2 Algorithm Selection

Table 8.4 shows the performances of the three algorithms to detect deteriorations when applied to wearable sensor and to intermittent data in turn. Performances are shown separately for the detection of SAEs and AEs. Higher AUROC values, indicating improved performance, were observed when detecting SAEs compared to AEs. This is to be expected since SAEs

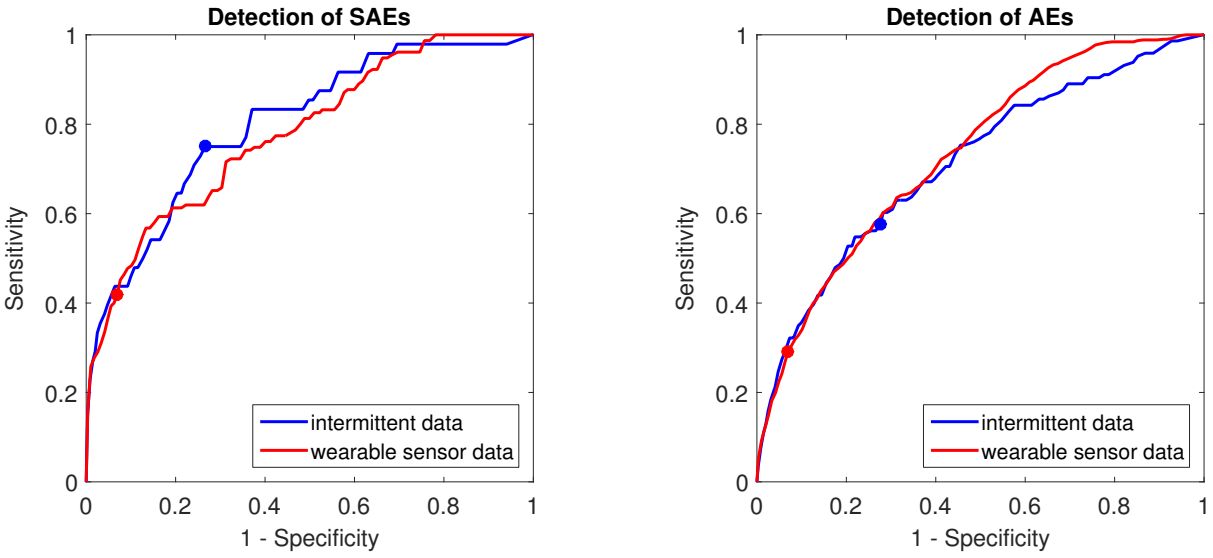

FIGURE 8.11: Receiver-operator curves for the $NEWS_{ai}$ algorithm when detecting severe adverse events (SAEs, left hand plot) or adverse events (AEs, right hand plot), using wearable sensor data (shown in red) or intermittent data (blue). Performances at the optimal thresholds are shown by dots (optimal thresholds were found as described in Section 8.4.3).

are likely to result in a greater level of physiological abnormality. When detecting SAEs, the AUROCs of the $NEWS_a$ and $NEWS_{ai}$ algorithms were similar regardless of whether wearable sensor data or intermittent data were used. This suggests that the wearable sensor parameters, including the continuous RR estimates, were sufficiently precisely measured to maintain the performances of the algorithms between intermittent measurements. When detecting AEs, the AUROCs of the $NEWS_a$ and $NEWS_{ai}$ algorithms were higher when using wearable sensor data than when using intermittent data. This suggests that wearable sensors may have utility for improving the detection of less severe deteriorations in ambulatory patients. The AUROCs of the KDE algorithm were higher when applied to wearable sensor data than when applied to intermittent data, which is to be expected since the KDE algorithm was trained using a model of normality extracted from the wearable sensor data. The $NEWS_{ai}$ consistently provided the best performance when detecting both SAEs and AEs. Therefore, it was chosen as the most suitable algorithm for use in the remainder of the study. The receiver-operator curves for this algorithm when applied to wearable sensor and to intermittent data are shown in Figure 8.11. The ROCs were similar when using intermittent data or wearable sensor data.

8.4.3 Performance Assessment

Figures 8.12 and 8.13 show how the PPV and the alert rate of the algorithm varied at different candidate alert thresholds. Both the PPVs and the alert rates were similar between applying

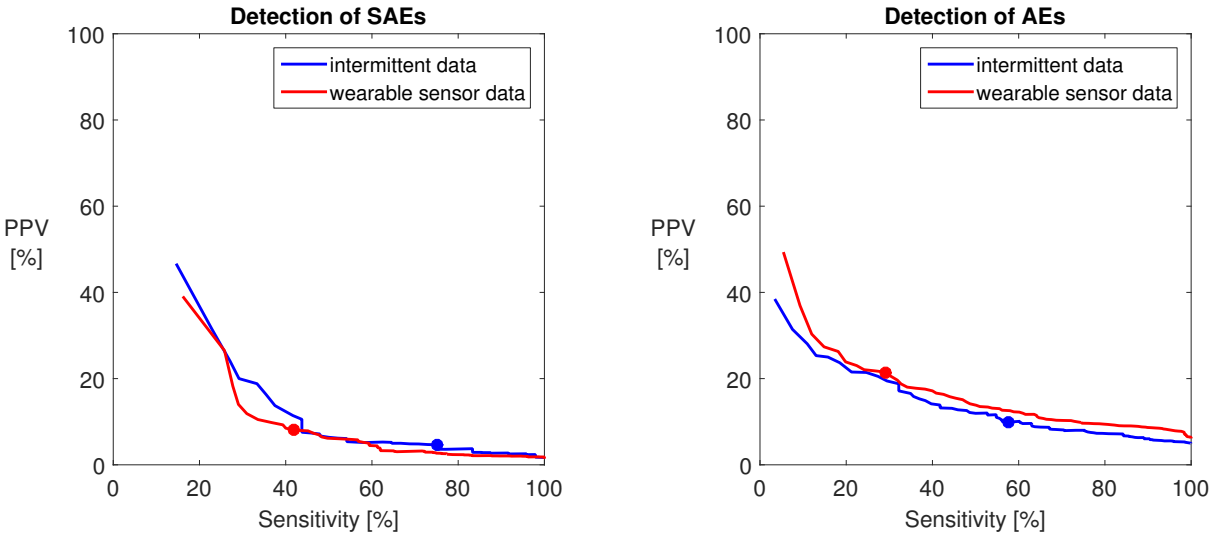

FIGURE 8.12: The PPV of the NEWS_{ai} algorithm across the range of candidate thresholds. Performances are shown for detection of SAEs (left hand plot) and AEs (right). Performances at the optimal thresholds are shown by dots.

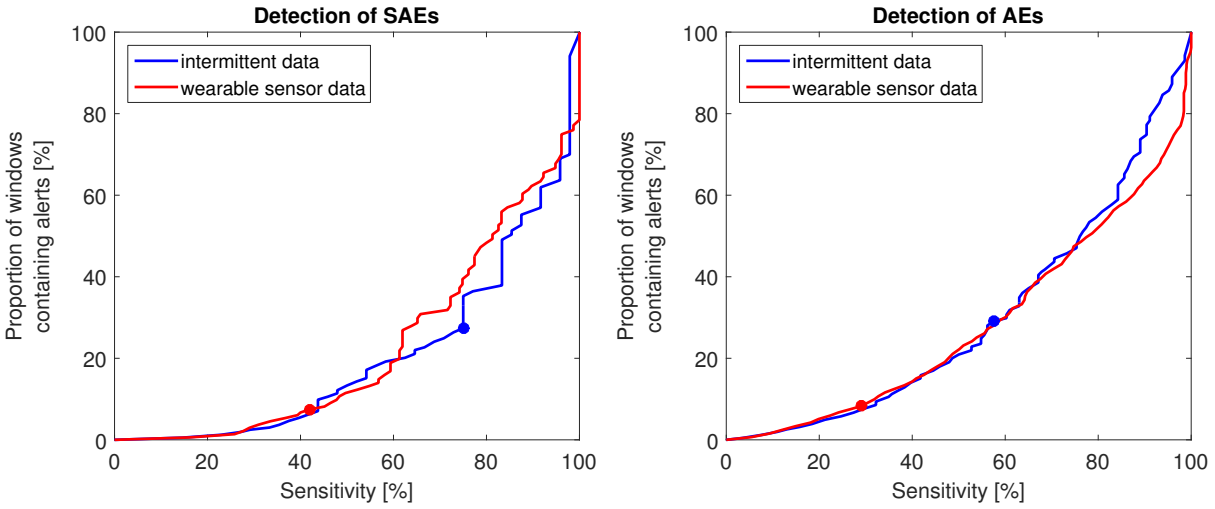

FIGURE 8.13: The proportion of windows which contained NEWS_{ai} alerts across the range of candidate thresholds.

the algorithm to intermittent or wearable sensor data. This supports the observation that the continuous input data to the algorithm provided by wearable sensors were sufficiently precise to maintain algorithmic performance between intermittent observations. There is a clear compromise between different aspects of algorithmic performance when selecting an alerting threshold. Increasing the sensitivity of the algorithm (which increases the proportion of deteriorations detected) reduces the algorithm's specificity (Figure 8.11), decreases its PPV, and increases the alert rate.

Optimal alerting thresholds were selected for the algorithm when using wearable sensor data ($k_{\text{w,NEWSai}}$) and intermittent data ($k_{\text{i,NEWSai}}$) by limiting the frequency of false alerts to one

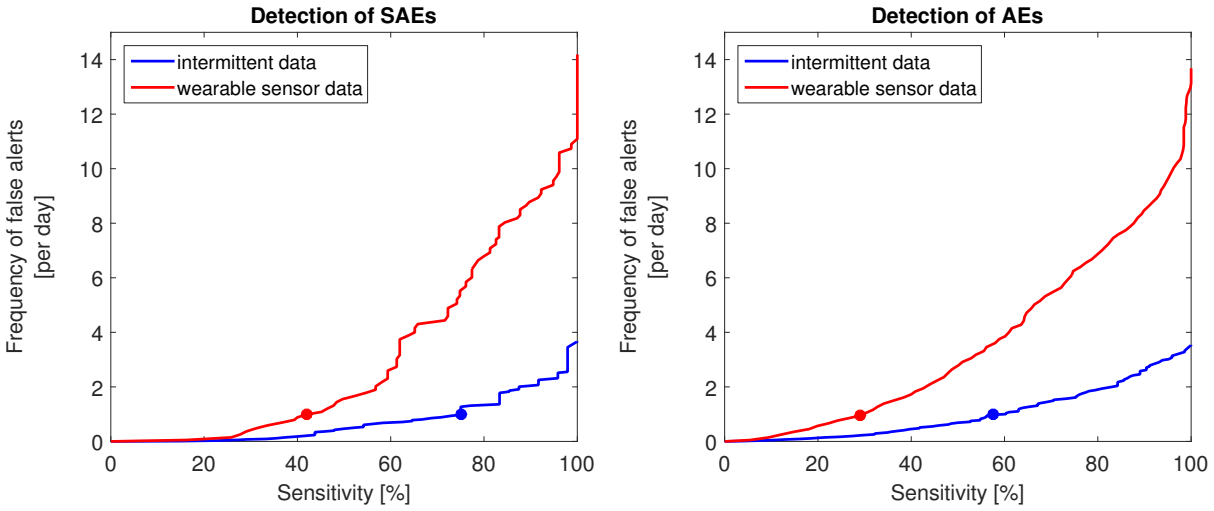

FIGURE 8.14: The frequency of false alerts when using the NEWS_{ai} algorithm across the range of candidate thresholds.

TABLE 8.5: The performance of the NEWS_{ai} algorithm when applied to wearable sensor data and to intermittent data

	Wearable sensor data		Intermittent data	
Performance metric	SAEs	AEs	SAEs	AEs
Sensitivity [%]	41.9	29.1	75.0	57.5
Specificity [%]	93.0	93.0	73.4	72.4
PPV [%]	8.0	21.5	4.7	9.9
Proportion of windows containing alerts [%]	7.4	8.3	27.4	29.1
Frequency of false alerts [per day]	1.0	1.0	1.0	1.0

per day. Figure 8.14 shows how the frequency of false alerts varied at different thresholds. The optimal threshold for use with wearable sensor data was found to be $k_{\text{w,NEWSai}} = 6.3$ when detecting both SAEs and AEs, and the optimal threshold for use with intermittent data was $k_{\text{i,NEWSai}} = 3.0$. The performances of the algorithm when using these thresholds are indicated by the dots on each figure, and are summarised in Table 8.5. The threshold selected for use with wearable sensor data was more stringent (requiring a greater level of physiological abnormality) than that for use with intermittent data. This was because intermittent data were acquired less frequently, and so a less stringent threshold could be used and still result in the same frequency of false alerts per day. Consequently, the algorithm had a lower sensitivity when using wearable sensor data compared to intermittent data (*e.g.* 41.9 % compared to 75.0 % when detecting SAEs), indicating that a lower proportion of deteriorations were detected. However, the PPV of the algorithm when using wearable sensor data was approximately twice as high as when using

intermittent data.

8.4.4 Case Studies

Four case studies are now presented demonstrating the utility of the $NEWS_{ai}$ algorithm. Each case study is accompanied by a plot of the $NEWS_{ai}$ score and the physiological parameters continuously monitored by wearable sensors. In these figures wearable sensor measurements are shown by small dots, and intermittent measurements by large dots. AEs and SAEs are annotated. Alerting $NEWS_{ai}$ scores are shown in red, and non-alerting scores in green.

The first case study is of a normal patient's recovery on the ambulatory ward, shown in Figure 8.15. The patient arrived on the ambulatory ward two days into their recovery from surgery. The $NEWS_{ai}$ algorithm did not alert at all throughout their stay when calculated using intermittent data. When calculated using wearable sensor data, it alerted briefly at the time of 3 hours. This was classified as a false alert in the above analysis, since it was not followed by an AE or SAE. It was caused by a rapid desaturation, accompanied by an increase in HR. The low SpO_2 measurement was quickly assessed by ward staff (as shown by the large dot indicating a manual measurement), and supplementary oxygen administered. The SpO_2 and HR values quickly returned to normal, and the $NEWS_{ai}$ score dropped to zero. Despite being classified as a false alert, this alert would have been clinically useful (prompting the administration of supplementary oxygen). For the next three days the algorithm did not alert, and the patient was discharged from hospital after a relatively short length of stay of 5.1 days, indicating a quick recovery. Throughout the monitoring period the continuous RR estimates agreed closely with the intermittent manual observations.

The second case study is of a 24 hour period of wearable sensor monitoring taken from an abnormal patient's recovery. The 24 hour period, shown in Figure 8.16, culminated in three SAEs in quick succession: a cardiac arrest, an unplanned return to surgery, and readmission to critical care. The $NEWS_{ai}$ algorithm alerted consistently throughout the first 12 hours, even though both the HR and SpO_2 measurements were within normal ranges. The alerts were in part caused by the continuous RR estimates being elevated at approximately 30 bpm. Whilst the intermittent HR and SpO_2 values during this period agreed closely with the corresponding wearable sensor measurements, the intermittent RR values were markedly different. Whereas the continuous RR estimates were elevated, resulting in continuous $NEWS_{ai}$ scores of approximately 7, the intermittent RR measurements were within the normal range, resulting in lower

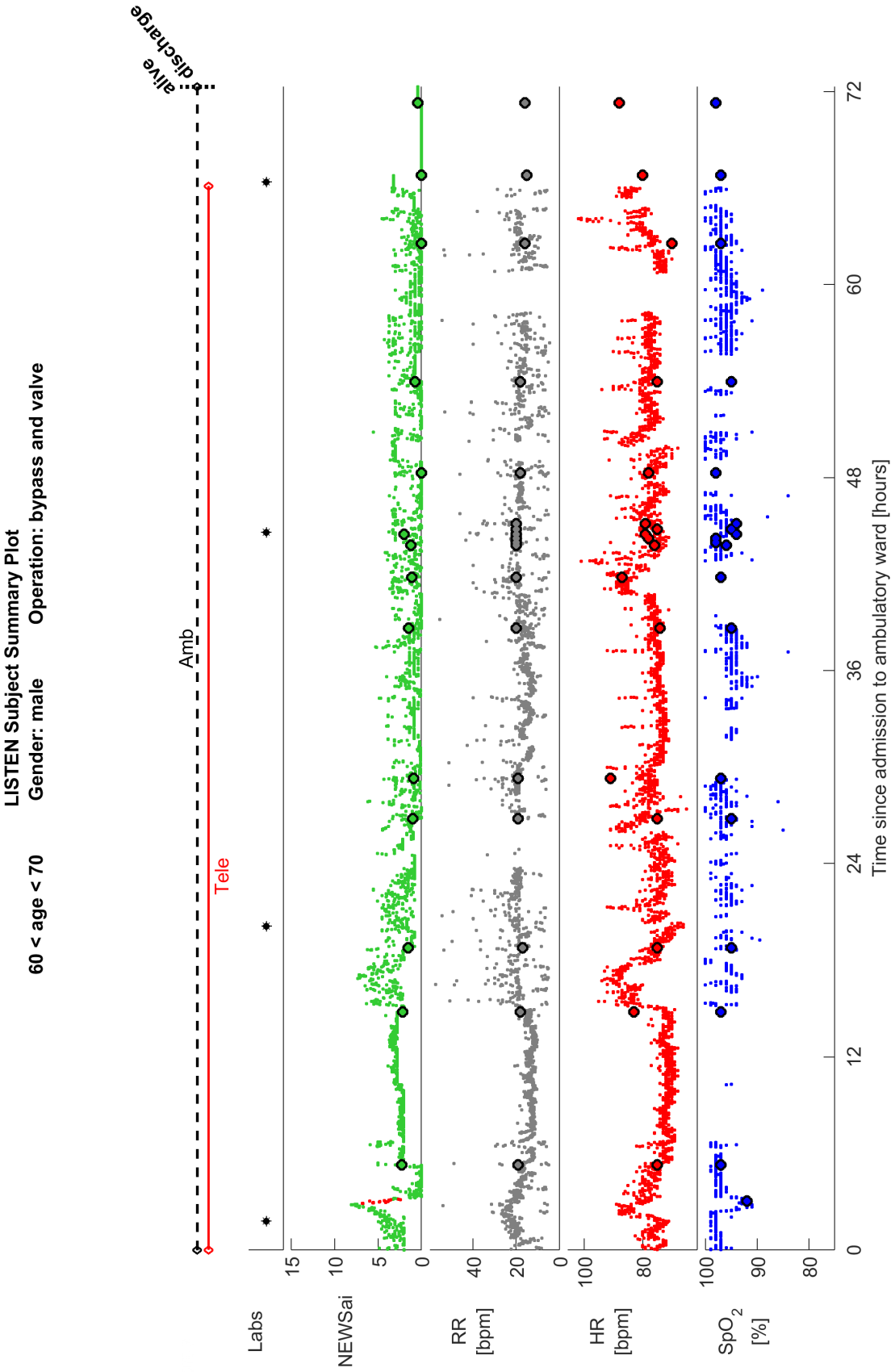

FIGURE 8.15: Case study of a normal patient's recovery on the ambulatory ward: This figure shows the continuous $NEWS_{ai}$, RR, HR and SpO_2 measurements derived from wearable sensor data (small dots), and the corresponding measurements derived from intermittent data (large dots). The $NEWS_{ai}$ algorithm alerted only once (as indicated by red, rather than green dots), at a time of 3 hours, indicating the need for supplementary oxygen. The patient made a quick recovery and was discharged from hospital after 72 hours on the ambulatory ward.

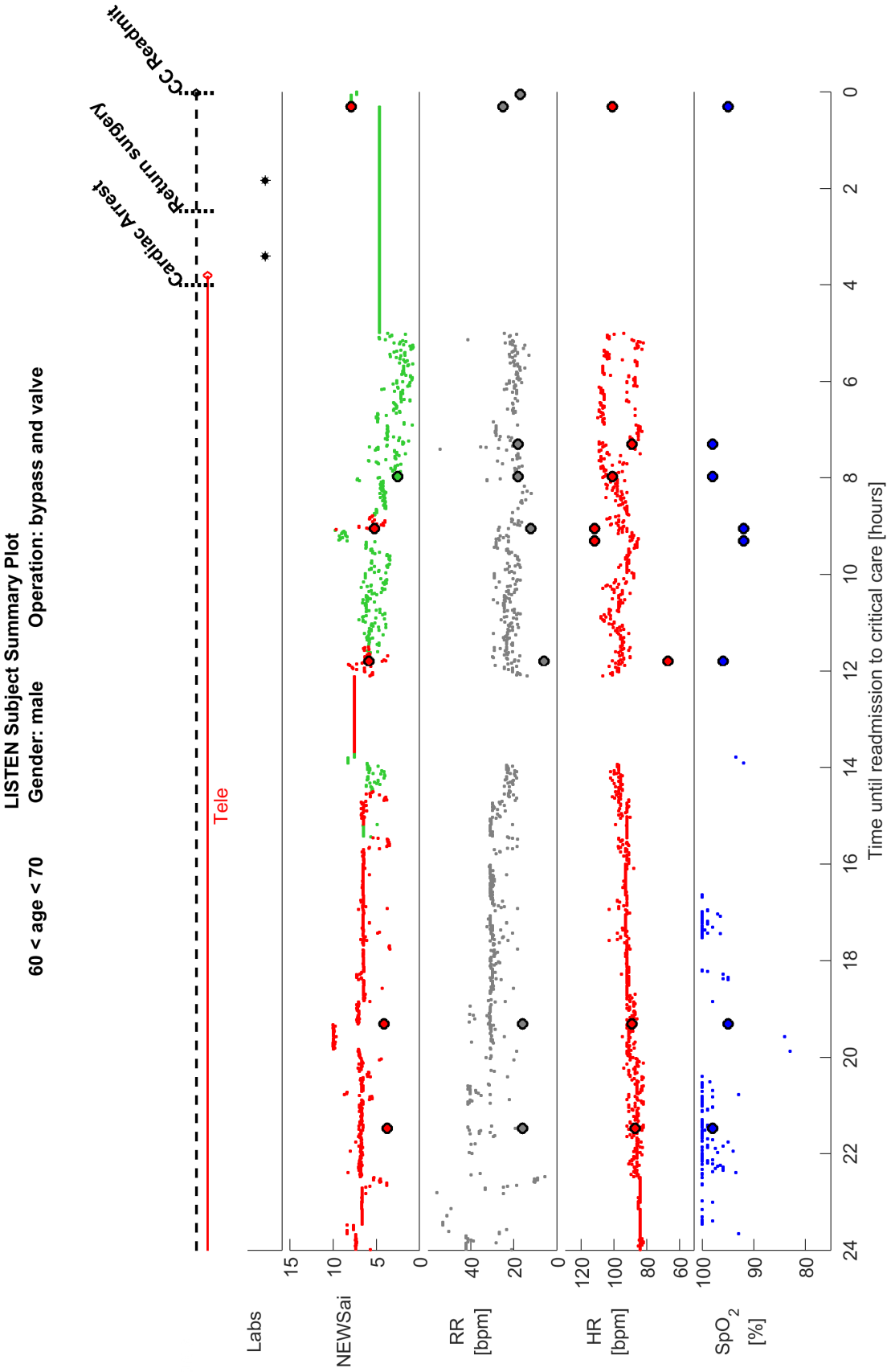

FIGURE 8.16: Case study of a 24 hour period culminating in a cardiac arrest, return to surgery and readmission to critical care. The NEWS_{ai} algorithm alerted consistently throughout the first 12 hours, caused in part by elevated RR estimates of 30 bpm, indicating a high level of physiological abnormality. Approximately 8 hours later the patient suffered a cardiac arrest.

intermittent NEWS_{ai} scores of approximately 4. Therefore, the level of physiological abnormality exhibited by the patient may not have been fully recognised by ward staff. In the second half of the recording the RRs returned to normal (approximately 20 bpm). The parameters no longer indicated a high likelihood of deterioration, and the algorithm stopped alerting. Despite this, the patient suffered a cardiac arrest only hours after the NEWS_{ai} alerts. This was followed by a prolonged hospital stay, including over two weeks in critical care and a further month on the ambulatory ward.

The first two case studies were both of male patients between 60 and 70 years of age, undergoing combined coronary artery bypass graft (CABG) and heart valve operations. In contrast, the third case study, shown in Figure 8.17, concerns the ambulatory ward stay of a male patient between 40 and 50 years of age, a relatively young patient. The NEWS_{ai} algorithm alerted consistently throughout this patient's stay, both when calculated using intermittent and wearable sensor data. The patient experienced an AE after 13 hours in the ambulatory ward: paroxysmal atrial fibrillation (AFib). Although this is common amongst patients recovering from cardiac surgery, in this instance it was accompanied by a drop in SBP from 123 to 99 mmHg. The patient's cardiovascular state was poor, with a variable cardiac output caused by the AFib, and a low blood pressure. In addition, the continuous RR estimates were frequently elevated to highly abnormal levels. This was reflected in extremely high NEWS_{ai} scores of over 10. The NEWS_{ai} scores dropped to approximately 8 upon the cessation of AFib at 22 hours. However, in the following hours the SBP remained as low as 78 mmHg, and the SpO_2 was as low as 90 % despite receiving supplementary oxygen. After two days of consistent NEWS_{ai} alerts, the patient was returned to surgery and readmitted to critical care. The patient died only three days later.

The fourth case study provides an example of when the NEWS_{ai} performed poorly, providing many false alerts despite the patient recovering well and being discharged after an uneventful hospital stay. Figure 8.18 shows the patient's data from throughout their post-operative stay, both in critical care (initial three days) and on the ambulatory ward (subsequent four days). The NEWS_{ai} algorithm generated many false alerts between 3 and 6 days, despite the patient not suffering from adverse events, and having a typical length of stay of 6.9 days (the median length of stay was 7.0 days, as noted in Table 3.1). The algorithm generated false alerts when calculated using either intermittent data, or the combination of intermittent and wearable sensor data, indicating that these false alerts were due to poor algorithmic performance rather than unreliable wearable sensor data. The patient exhibited slightly deranged physiology during the

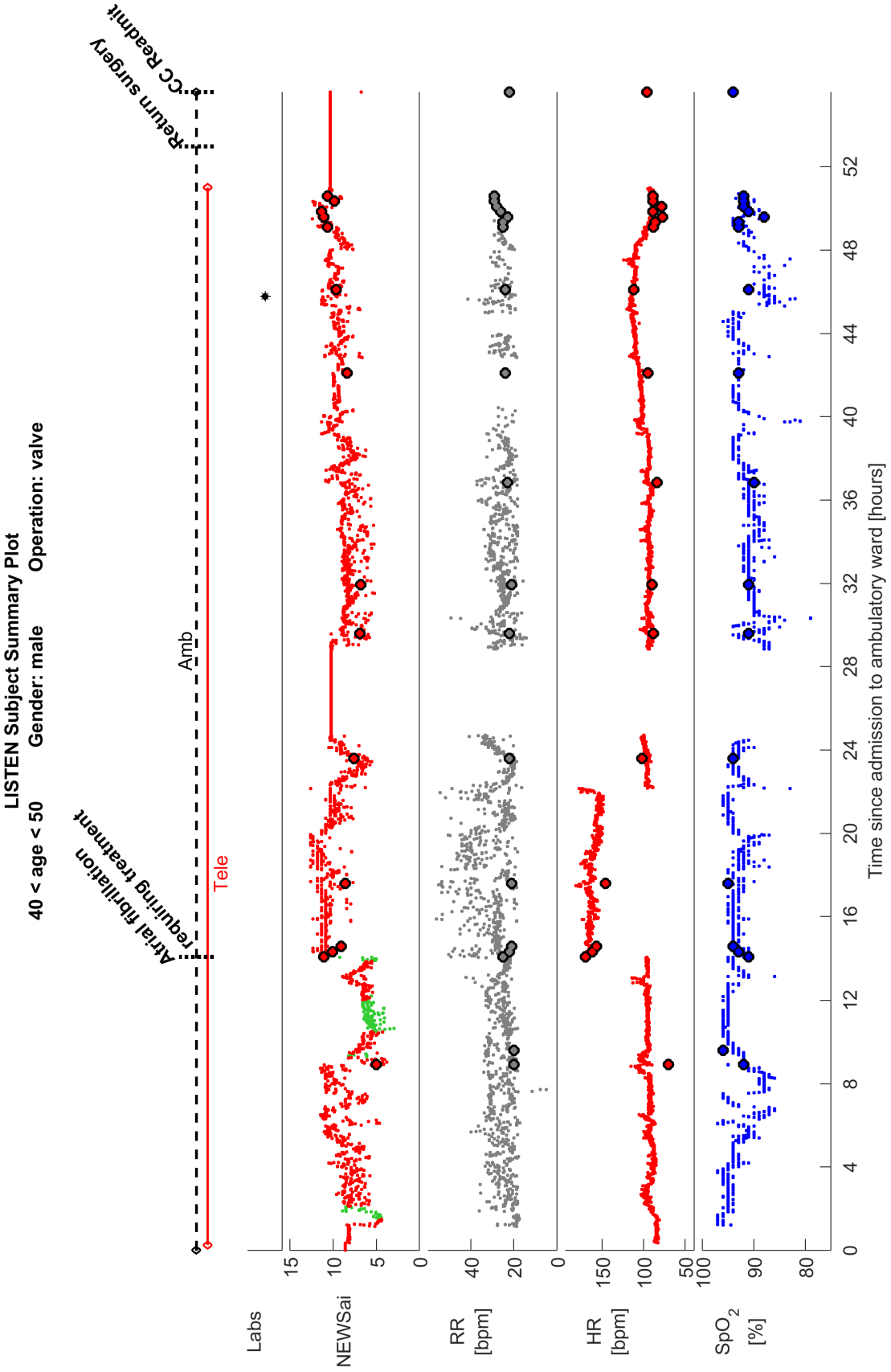

FIGURE 8.17: Case study of an abnormal patient's recovery on the ambulatory ward: The NEWS_{ai} algorithm alerted consistently for over two days, culminating in the patient returning for surgery, and being readmitted to critical care. The patient died only 72 hours later.

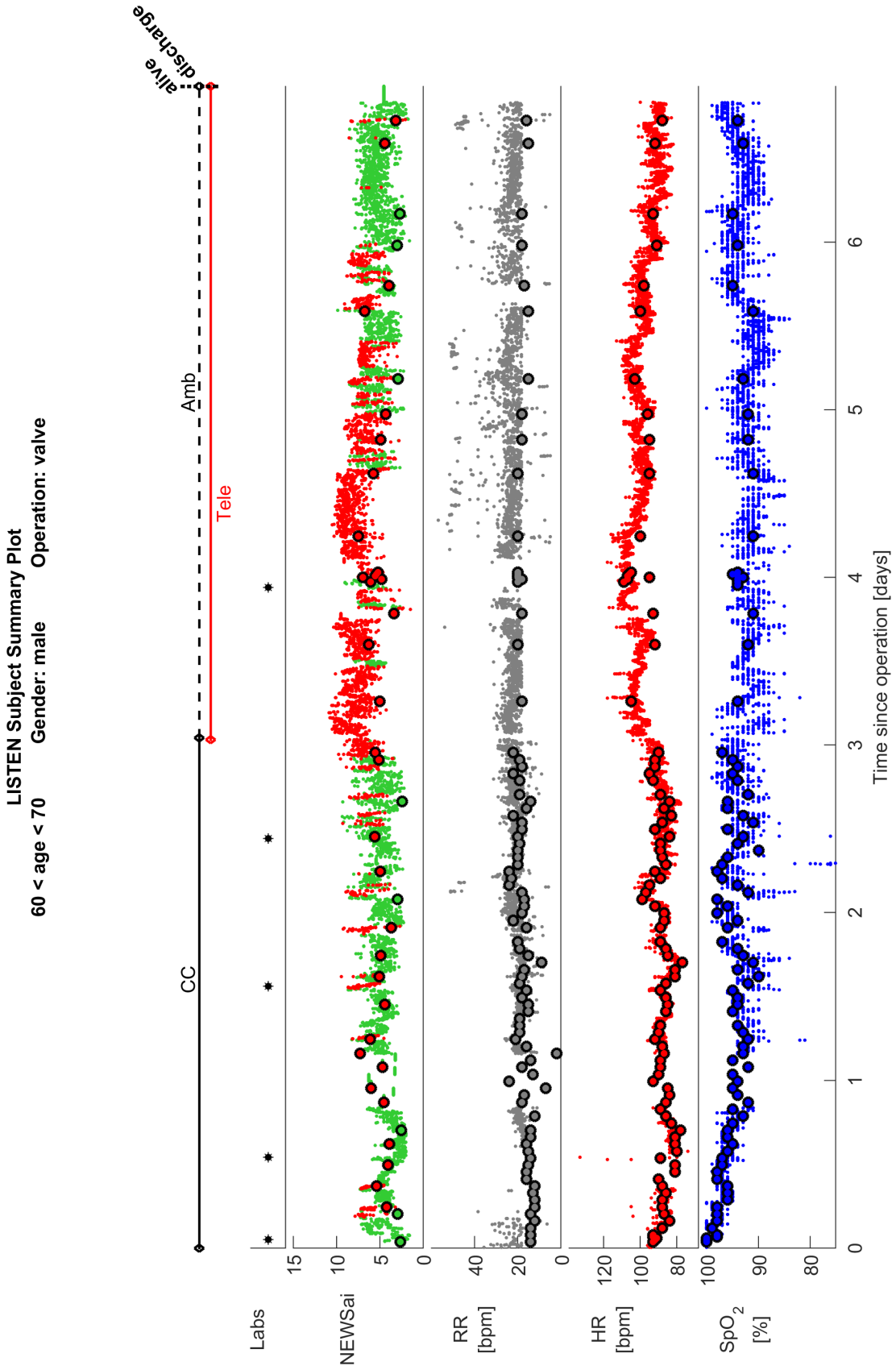

FIGURE 8.18: Case study of a normal patient's recovery with many false alerts: in which the NEWS_{ai} algorithm provided many false alerts: This plot shows data from a patient's entire post-operative stay, in critical care wards for the first three days, followed by monitoring using a wearable sensor on the ambulatory ward for the latter four days. The NEWS_{ai} algorithm alerted for much of the first three days on the ambulatory ward, despite the patient recovering well and being discharged from hospital approximately one week after their operation.

period of false alerts of : HR ≈ 100 bpm, SpO_2 ≈ 94 %, SBP ≈ 107 mmHg, intermittent RR ≈ 18 bpm, and continuous RR ≈ 23 bpm. In this case study of a patient classified as normal, slightly deranged physiology across multiple parameters resulted in elevated $NEWS_{ai}$ scores and alerts.

8.5 Discussion

This study demonstrated the feasibility of detecting deteriorations using multi-parametric wearable sensors, coupled with an algorithm to estimate RR continuously from the ECG. The algorithm used to detect deteriorations was a modified version of the NEWS in which the scoring system was interpolated to avoid the quantised outputs provided by the standard NEWS. This algorithm acted on three median-filtered parameters (HR, SpO_2 and RR), sampled once per minute, and three intermittent parameters acquired intermittently by hand (SBP, temp and O_2). Alerts were generated if the algorithm's output consistently breached a threshold indicating increased likelihood of deterioration. The algorithm exhibited similar performance when applied to both wearable sensor and intermittent data, demonstrating the feasibility of using unverified wearable sensor data to detect deteriorations. However, as the algorithm provided more frequent assessments of the likelihood of deterioration when applied to wearable sensor data, the frequency of false alerts at any particular alerting threshold was increased. Consequently, the threshold had to be set at a much higher value when using wearable sensor data compared to intermittent data to maintain a clinically manageable false alert rate. As a result the sensitivities of the algorithm when applied to wearable sensor data were 41.9 and 29.1 % for detection of SAEs and AEs respectively, indicating that a large proportion of deteriorations were not identified. This suggests that further work is required to improve the performance of algorithms to detect deteriorations from wearable sensor data before they can be expected to be clinically beneficial.

Several studies have previously evaluated multi-parametric wearable sensors for continuous monitoring of hospital patients [33, 66, 75, 77, 84–91, 91, 92] (see Table 2.4 for details). However, few studies have investigated the utility of multi-parametric wearable sensor data for detection of deteriorations. The evidence provided by this study suggests that early warning scores can be continuously updated using the unverified data provided by wearable sensors, with no loss in performance. This is notable since wearable sensor data are vulnerable to imprecise measurements, and measurements which are not representative of the patient's physiological status at

rest (as described in Section 8.2). The use of signal quality assessment algorithms and median-filtering to clean data ensured that the algorithm's performance was maintained between the intermittent manual measurements. In the future this may confer two potential clinical benefits (as described in Section 2.3). Firstly, it may be possible to provide early warning of SAEs and AEs using wearable sensors, since they may detect changes in physiology in-between routine manual assessments. Secondly, it may be possible to reduce the workload associated with manual assessments by reducing their frequency at times when the continuously updated early warning score remains low.

In this study the NEWS_{ai} algorithm did not perform sufficiently well to expect that the potential benefits of wearable sensors for detection of deteriorations could be realised without further algorithmic development. Its AUROC for detection of AEs (0.780 for wearable sensor data, and 0.794 for intermittent data) was lower than the value of 0.873 reported previously [45]. There are several possible reasons for this impaired performance. Firstly, the normal patients in this dataset exhibited abnormal physiology (see Figure 8.10), which may have impaired the algorithm's ability to distinguish between normal and abnormal patients. Secondly, the dataset used in this study did not contain LOC, which has previously been shown to be a valuable predictor of in-hospital mortality [331]. Thirdly, this study used data from the narrow population of patients recovering from cardiac surgery, and this population may not be representative of all acutely-ill patients. For instance, it suffers from a high incidence of paroxysmal AFib [364]. It may be necessary to refine algorithms to detect deteriorations prior to using them continuously with wearable sensor data. Potential refinements include: (i) the use of additional variables commonly recorded in electronic health records (EHRs), such as demographics, laboratory test results, and co-morbidities [11, 23, 121, 130]; (ii) the use of alternative machine learning techniques such as tree-based models, support vector machines, logistic regression and Gaussian processes [348, 365]; and, (iii) the use of vital sign trends to identify trajectories of deterioration [366, 367].

In this study three algorithms were used to assess the likelihood of deterioration, with the NEWS_{ai} providing superior performance over the NEWS_{a} and KDE algorithms. This may have been because the interpolated score calculation reduced the frequency of false alerts, or because the calculation more precisely evaluates the abnormality of each vital sign, increasing the precision of the assessed likelihood of deterioration. Surprisingly, the KDE algorithm performed less well than the others, despite its model of normality being constructed using training data from the dataset. This may be because the training dataset was relatively small (6.1×10^4

patterns), compared to the training dataset in the original development of the KDE algorithm (2.3×10^6 patterns) [354]. In addition, when using the KDE technique a single parameter (such as an extremely abnormal HR of 140 bpm due to AFib) can contribute very strongly towards the final score, whereas in the NEWS_{ai} model the contribution of any single parameter is always limited to a score of three.

In this study RRs were estimated from the ECG signal using the novel algorithm presented in Chapter 7. The performance of this algorithm has been assessed extensively using a range of datasets acquired from both healthy and sick subjects (see Chapter 7). However, its performance in the ambulatory setting has not yet been assessed. The second case study (Figure **??**) demonstrated the potential benefits of continuous RR monitoring, since the alerts preceding a cardiac arrest in this example were greatly contributed to by the high RR estimates. However, some examples of poor performance were observed as described in Section 8.3.1. It has traditionally been difficult to assess the performance of RR algorithms in the ambulatory setting, since it is difficult to obtain a precise reference RR. However, recently developed wearable sensors can acquire both ECG and impedance pneumography (ImP) signals. Such equipment could be used to acquire a dataset with which to assess the algorithm in the ambulatory setting when used with the novel ImP signal quality index presented in Section 4.6.2.

8.6 Final Remarks

This chapter has presented a study investigating the feasibility of using wearable sensor data to continuously assess the likelihood of deterioration of acutely-ill, ambulatory patients. The novel algorithm presented in Chapter 7 was used to continuously estimate RR from the ECG. A modified version of the NEWS designed specifically for wearable sensor data was selected to assess likelihood of deterioration. The algorithm was found to provide similar performance when applied to wearable sensor data as when applied to intermittent data, suggesting that it is feasible to use wearable sensors to continuously assess the likelihood of deterioration. However, the use of wearable sensor data produced more frequent outputs, causing the rate of false alerts per day to be increased in comparison to the use of intermittent data. This suggests that further work is required to improve the performances of algorithms to detect deteriorations before they can be expected to confer widespread clinical benefit when used with wearable sensors.

Chapter 9

Conclusion

This chapter concludes the thesis by presenting an overview of its achievements (Section 9.1), and recommended directions for future research (Section 9.2).

9.1 Summary of Thesis Achievements

The aim of this thesis was to develop and assess the performance of techniques for continuous respiratory rate (RR) monitoring using electrocardiogram (ECG) and photoplethysmogram (PPG) signals for use in wearable sensors to detect deteriorations. Chapters 3 to 8 presented investigations to address this aim. The achievements of the thesis are summarised as follows:

Chapter 3: Assembling a comprehensive physiological dataset using wearable sensors

Chapter 3 presented the data processing conducted to assemble a comprehensive physiological dataset suitable for the development of algorithms to detect deteriorations from wearable sensor data. In the LISTEN Trial, data were collected from 226 patients during their recovery from major cardiac surgery. Of these, 200 patients were provided with a wearable sensor, which was worn for a median (lower - upper quartiles) of 3.7 (2.6 - 5.2) days. The data acquired in this trial were processed to provide two datasets suitable for use in this thesis and in future research: The LISTEN Prepared Dataset, and the LISTEN Processed Dataset. The LISTEN Prepared Dataset is designed for a minority of investigations which require raw monitoring data, such

as investigations into the efficacy of using wearable sensors to acquire data continuously from ambulatory patients. The LISTEN Processed Dataset is more suitable for the development of algorithms to detect deteriorations, since the continuous data have been quality assessed to eliminate low quality measurements. The wearable sensor data consist of continuous electrocardiogram (ECG) and photoplethysmogram (PPG) signals, routinely acquired heart rate (HR) and arterial blood oxygen saturation (SpO_2) parameters, and additionally RR estimates derived from the ECG using the algorithm presented in Chapter 7. In addition to wearable sensor data, the datasets contain: (i) continuous monitoring data acquired using static bedside monitors in critical care; (ii) intermittently measured physiological data (such as vital signs measurements taken manually by ward staff), from which benchmark early warning scores (EWSs) can be calculated; (iii) laboratory test results; (iv) end-points for training machine learning algorithms (such as adverse events - AEs, and severe adverse events - SAEs); (v) fixed variables (such as demographic data); and (vi) times of transfers between wards and levels of care. This is a valuable resource since it contains wearable sensor data acquired for several days from a large set of acutely-ill patients.

After describing the processes undertaken to assemble the datasets, the LISTEN Processed Dataset was evaluated for the purposes of this thesis. It was observed that the data coverage rate was higher for the ECG signal than for the PPG signal, which informed the decision to estimate RRs continuously from the ECG signal.

Chapter 4: Identifying algorithms to estimate RR from the ECG and PPG from the literature, and methods used to assess their performance, to create a toolbox of resources for algorithm assessments

Chapter 4 presented a systematic review of algorithms to estimate RR from the ECG and PPG, and the methods previously used to assess their performances. The findings of the review informed the design of a toolbox of resources for assessments of RR algorithm performances.

The systematic review identified 140 publications describing RR algorithms for use with the ECG or PPG, using both manual and electronic searches. A generalised structure of RR algorithms was proposed, consisting of five algorithmic stages, and the use of any one of a number of interchangeable techniques at each stage. A total of 49 previously proposed interchangeable techniques were identified. Several aspects of algorithm assessment methodologies were common in the literature: (i) most assessments used window durations of between 30 and 90 s; (ii)

only two publications compared the performances of > 15 algorithms (one of which arose from this work [10]); (iii) the datasets used tended to include data from young and healthy adults, with fewer assessments using data from elderly adults or patients, and only 5.7 % using data from acutely-ill patients; (iv) eleven publicly available datasets have been used to assess RR algorithms, although only one publication was identified which used more than two datasets; (v) impedance pneumography (ImP) and air flow/pressure signals have been used most often to obtain reference RRs, although the performance of algorithms to estimate reference RRs was usually unknown; and, (vi) a range of statistics have been used to assess performance, with no clear consensus as to the most appropriate approach.

The findings of the review were used to identify key areas for future research, which were then addressed by designing a toolbox of resources to facilitate research into these areas. The three directions identified for future research were: (i) to perform a comprehensive comparison of algorithms; (ii) to assess their performance in acutely-ill patients; and (iii) to develop techniques to ensure that reference RRs are reliable. Consequently, the toolbox contains three features to address these research areas. Firstly, novel datasets were collated from clinical trial data, which are being made publicly available. The datasets consist of: (i) synthetic data to ensure algorithm implementations are reasonable; (ii) data from young healthy subjects to assess algorithms in ideal conditions [10]; (iii) data from young and elderly healthy subjects to assess the impact of age on algorithms; (iv) data from 108 acutely-ill hospitalised patients to assess performance in this population; (v) a dataset to assess the impact of atrial fibrillation (AFib) on algorithms; and (vi) a dataset to assess the impact of mechanical ventilation on algorithms. Secondly, a novel signal quality index (SQI) for the ImP signal, which is commonly used to estimate reference RRs. It was designed and validated using manual annotations of 1801 signal segments from 34 patients across three clinical settings, and was shown to have high accuracy and precision when used to estimate reference RRs. Thirdly, a toolbox of algorithms was created using Matlab ®, containing 370 algorithms formed through combinations of 31 interchangeable algorithmic techniques [10, 11].

Chapter 5: Assessing the influence of technical and physiological factors on respiratory signals extracted from the ECG and PPG

Chapter 5 presented a study to assess the influence of technical and physiological factors on the quality of respiratory signals extracted from the ECG and PPG, using the RRest-healthy

dataset collated in Chapter 4. Quality was assessed using the correlation between an extracted respiratory signal and a reference respiratory signal. A total of thirteen respiratory signals were extracted from the ECG and PPG signals. The study primarily used data from 57 healthy subjects, acquired in ideal conditions using multiple measurement devices. In addition, the publicly available Fantasia dataset [302, 303] was used for external verification of a subset of the results. The results informed recommendations regarding wearable sensor designs when incorporating an algorithm to estimate RR from the ECG or PPG, and regarding potential clinical applications.

The recommendations for wearable sensor design were: (i) the PPG should be measured at the finger rather than the ear, in contrast to the findings of previous work which have suggested using the ear [319], [322]; (ii) current clinical monitoring equipment can be used to acquire ECG and PPG signals suitable for extraction of high quality respiratory signals; (iii) the ECG should be used rather than the PPG for measuring respiratory signals when available; and, (iv) the minimum recommended sampling frequencies for the ECG and PPG for extraction of respiratory signals were 250 Hz and 16 Hz respectively. The recommendations regarding potential clinical applications were as follows: (i) it may be helpful to avoid the use of respiratory signals which are based on frequency modulation (FM) in elderly subjects (such as when using wearable sensors to monitor acutely-ill patients); (ii) no evidence was found to suggest that gender influences the quality of respiratory signals; (iii) the qualities of most respiratory signals were lower at higher RRs, suggesting that RR algorithms may not be as precise at higher RRs. These findings will be useful to equipment designers, and clinicians, for optimising the performance of RR algorithms in wearable sensors.

Chapter 6: Assessing the performance of RR algorithms for the ECG and PPG in healthy subjects and hospitalised patients

Chapter 6 presented a comprehensive assessment of the performances of RR algorithms in healthy subjects and hospitalised patients, using the ImP SQI, and the toolbox of algorithms designed in Chapter 4. Data from four datasets were used to ensure the generalisability of the results, and a total of 95 RR algorithms were used to identify the most suitable algorithm for use with wearable sensors. Algorithm performances were found to vary greatly between different RR algorithms, and between datasets, both in healthy subjects and hospitalised patients. Key

features of RR algorithms were identified which provide improved performance. In health, improved performance was achieved using a time-domain RR estimation technique, and by using a modulation fusion technique. The best-performing algorithms only performed sufficiently well (in comparison to routine RR monitoring) when used with ECG signals in healthy subjects, suggesting that none of the algorithms assessed performed well enough for use in wearable sensors with acutely-ill patients. Key reasons for this were: frequency-domain RR estimation techniques frequently underestimate RR substantially, and time-domain RR estimation techniques generally have a lower level of precision when an algorithm performs well. This suggested that a novel algorithm was required to estimate RR from the ECG or PPG for use in wearable sensors.

Chapter 7: Developing a novel algorithm for continuous RR monitoring

Chapter 7 presented a novel RR algorithm designed specifically to provide precise RR estimates when using wearable sensors. It was designed to harness the strengths of both time- and frequency-domain RR estimation techniques. A frequency-domain technique was initially used to estimate RR, and if the RR estimate was greater than an empirical threshold then it was accepted. However, if the RR estimate was less than the threshold then a time-domain technique was also used, and a final estimate was only provided if the two estimates were in close agreement. This was designed to exploit the high precision of frequency-domain RR estimation techniques, whilst eliminating the underestimates provided by these techniques. In addition, the novel RR algorithm was designed to only output RR estimates which were likely to be precise, in contrast to many existing algorithms. The empirical thresholds used in the algorithm were optimised using two datasets acquired from healthy subjects.

The performance of the novel RR algorithm was assessed using the same procedure as used in Chapter 6, ensuring that the results were directly comparable to existing algorithms. It was primarily assessed using data from 127 subjects in the MIMIC-II dataset [299, 303]. It was found to perform well when using the ECG, providing RR estimates for 40 % of windows, with 95 % of absolute errors $\leq$ 3.2 breaths per minute (bpm). For comparison, the errors of the best-performing algorithm assessed in Chapter 6 were $\leq$ 8.6 bpm. Furthermore, only 3.2 % of the absolute errors were $>$ 5.0 bpm, indicating a high level of clinical utility. However, the algorithm did not perform well enough for use in wearable sensors when applied to the PPG signal.

Chapter 8: Assessing the potential utility of wearable sensors for detecting deteriorations

Chapter 8 presented a study investigating the potential utility of wearable sensors for detecting deteriorations. The study used the LISTEN Processed Dataset, presented in Chapter 3, and the novel RR algorithm presented in Chapter 7. RRs were continuously estimated from the ECG signals acquired from 184 patients using wearable sensors. Deteriorations were detected using three algorithms: (i) an abbreviated National Early Warning Score ($NEWS_a$), in which level of consciousness was omitted from the routine scoring system; (ii) an interpolated NEWS ($NEWS_{ai}$); (iii) and a kernel density estimation (KDE) algorithm trained using data from normal LISTEN patients. Wearable sensor data were pre-processed using median-filtering to eliminate short-term variability, and alerts of deteriorations were activated and deactivated using empirically determined thresholds.

The $NEWS_{ai}$ algorithm was found to provide the best performance, and provided similar performance regardless of whether it was calculated using wearable sensor data, or routine intermittent data. This demonstrated that it is feasible to continuously update EWSs using data from wearable sensors, including continuous RRs estimated from the ECG. However, the frequency of false alerts per day was higher when using wearable sensor data since it is acquired continuously. Therefore, a higher alerting threshold was required to maintain a clinically manageable false alert rate when using wearable sensor data. Consequently, alerts of deteriorations were raised in less than half of the windows which were followed by SAEs or AEs within 48 hours (sensitivities of 41.9 % and 29.1 % respectively). Therefore, further work is required to improve the performance of algorithms to detect deteriorations from wearable sensor data before they can be expected to confer clinical benefit.

9.2 Future Work

Future research into the use of continuous RR monitoring in wearable sensors to detect clinical deteriorations will encounter several technical challenges which must be overcome to ensure that the full potential of wearable sensors is realised. These challenges are now discussed.

9.2.1 Further assessment and development of RR algorithms

This thesis has presented one of the most comprehensive assessments of RR algorithms to date. In Chapter 6 the performances of 95 algorithms were assessed using four datasets acquired from both healthy and hospitalised patients. Despite this, further assessment and development of RR algorithms may be required before they can be considered for widespread clinical use. Further assessment of performance is required since performance may vary between different patient populations. The datasets presented in Section 4.6.1 are currently being used by clinical collaborators to assess the performance of RR algorithms in acutely-ill patients (RRest-acute Dataset), and in patients with atrial fibrillation (AFib, RRest-afib Dataset). This is important because acutely-ill patients will benefit most from the use of RR algorithms in wearable sensors, and arrhythmias may adversely affect the performance of RR algorithms. However, relatively few algorithm assessments have been conducted previously in these populations ([148, 155, 209, 238, 249, 261, 265] describe assessments using acutely-ill patients, and [273] describes an assessment of the impact of AFib on performance). In particular, AFib may adversely affect the quality of respiratory signals extracted from the ECG and PPG [253, 273]. In addition, algorithm performance should be assessed in the intended setting for the algorithms: using data acquired from ambulatory patients over prolonged time periods. This has previously been difficult as it is challenging to acquire ECG or PPG signals alongside a reference respiratory signal for a prolonged time period in this setting. However, it is now possible to acquire these signals simultaneously from novel wearable sensors (see Section 8.5), providing opportunity to assess performance in this setting using routinely acquired data.

Further development of RR algorithms may be beneficial to enable them to be applied to other signals acquired by wearable sensors, such as accelerometer signals. Accelerometers are now being incorporated into novel wearable sensors. It may be beneficial to estimate RR from accelerometer signals either to provide a stand-alone estimate, or to fuse with estimates from other signals such as the ECG [163] to improve performance. Crucially, accelerometers may provide more precise RR estimates in the presence of cardiac arrhythmias. Previous research into estimation of RR from accelerometers has suggested that it is feasible [107, 163, 368–370], although further work is required to refine RR algorithms for use with accelerometers and to assess their performance with acutely-ill patients.

9.2.2 Improving algorithms to detect deteriorations

The conclusion of Chapter 8 was that further work is required to improve the performance of algorithms to detect deteriorations from wearable sensor data to avoid excessively high alert rates. There are at least three ways in which continuous, multi-parametric data provided by wearable sensors could be used to improve the performance of algorithms to detect deteriorations:

1. **Earlier identification of changes in physiology:** The continuous data provided by wearable sensors may provide earlier identification of changes in physiological state than current intermittent measurements [3]. In addition, wearable sensor data may provide early identification of movement towards a clinical state which is associated with a greater risk of adverse events. Evidence based on intermittent measurements is emerging suggesting that trends in parameters, rather than simply the absolute values, can be used to improve models for detection of deteriorations [367, 371]. The power of this approach may be greatly improved when using trend analysis algorithms with wearable sensor data, such as those in [372].

2. **Personalised alerting:** The high sampling rate of wearable sensor data may facilitate machine learning of personalised alert thresholds based on a quickly amassed historical dataset for each patient [354]. In contrast, the low sampling rate of intermittent measurements means it is not realistic to personalise alerting thresholds of EWSs. Therefore, current EWSs do not account for the heterogeneity of individual patients physiologies and disease progressions, and typically use the same alert thresholds for all patients.

3. **Variability analysis:** The continuous nature of wearable sensor data permits assessment of physiological variability, which cannot be performed at the same resolution with intermittent measurements. Variability analysis provides additional insight into the physiological state, beyond that provided by the absolute values of physiological parameters [373–376]. It can take two forms. Firstly, the variability of quasi-periodic phenomena, such as beat-to-beat or breath-to-breath intervals can be analysed to obtain information on autonomic function. For instance, heart rate variability (HRV) provides insight into the sympathetic-parasympathetic balance [375]. Both decreased and increased variability can be indicative of perturbations from health to disease states [374], and therefore offer potential for early detection of deteriorations [377]. Changes in variability have been associated with infection [378], sepsis [376], severity of illness and organ failure [379]. Secondly,

the variability of physiological parameters, such as temperature, can provide insight into physical activity or the maintenance of homeostasis.

9.2.3 Extraction of additional parameters from wearable sensor signals

The physiological signals monitored by wearable sensors contain a wealth of information on cardiac, vascular and autonomic function. However, very little of this information is utilised in wearable sensors at present. There is scope to facilitate more comprehensive physiological assessments using wearable sensors by increasing the range of parameters extracted from these signals. For instance, the ECG and PPG signals, considered in this thesis, contain information on several physiological systems including the cardiac, vascular, autonomic and respiratory systems. The ECG provides a rich source of information on the heart and autonomic nervous system. Its morphology is largely determined by the electrical activity of the heart. It is constructed of several waves and intervals, as illustrated in Figure 9.1. The PPG provides information on the heart and the vascular and autonomic nervous systems. A typical waveform is shown in Figure 9.2. When in sinus rhythm, beat-to-beat intervals of ECG and PPG waveforms are controlled by the autonomic nervous system, and contain information on the balance between sympathetic and parasympathetic activity [375]. Despite the wealth of physiological information available in the ECG and PPG waveforms, usually only HR and SpO_2 are extracted from them by wearable sensors. There is scope to extract a range of additional physiological parameters from the waveforms to facilitate more comprehensive physiological assessment using wearable sensors.

To this end, waveform analysis has recently been incorporated into wearable sensors. One wearable sensor now in clinical use provides measurements of ECG morphology which may be indicative of myocardial ischaemia [83]. Other sensors used in research provide beat-to-beat timings for HRV analyses [383]. In contrast, the PPG morphology is not currently used to derive additional parameters.

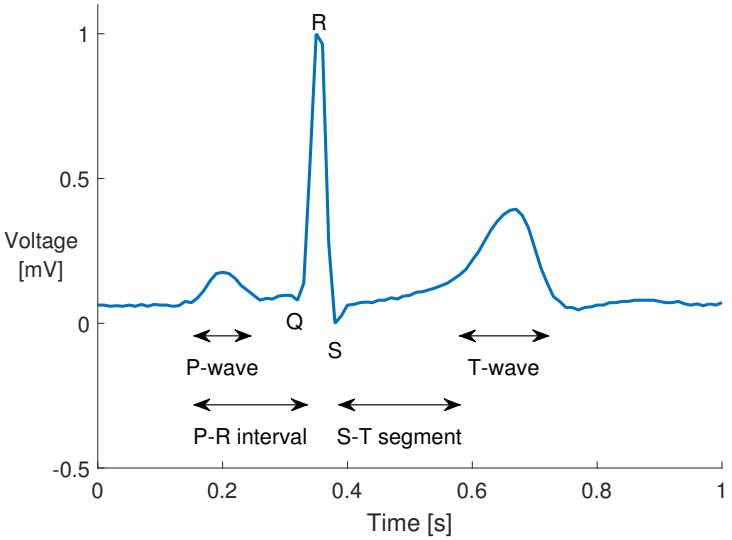

FIGURE 9.1: A typical electrocardiogram (ECG) waveform for one cardiac cycle [380]: The P-wave indicates atrial depolarization, with atrial contraction beginning at the end of the P-wave. The QRS complex indicates the onset of ventricular contraction. The T-wave indicates ventricular repolarisation. The P-R interval is the time from the beginning of the P-wave to the beginning of the QRS complex, approximately indicative of the time for an electrical impulse to travel from the sino-atrial node to the atrioventricular node. The S-T segment is the time from the S-wave to the beginning of the T-wave, indicative of ventricular contraction.

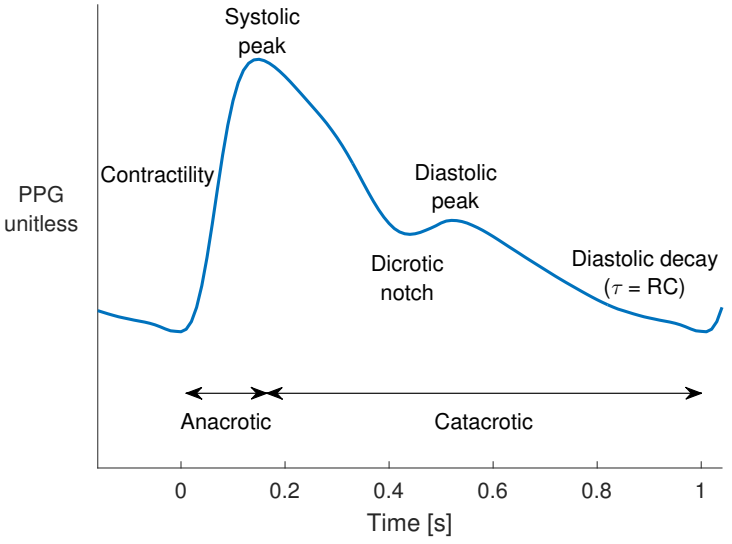

FIGURE 9.2: A typical photoplethysmogram (PPG) waveform for one cardiac cycle: The waveform can be separated into anacrotic and catacrotic phases, which are dominated by systolic ejection and wave reflections from the periphery respectively [381]. The systolic rising edge in the anacrotic phase is caused by the expansion of the arterial system due to inflow of blood. The rate of expansion is linked to the contractility of the heart, and the amplitude of the systolic peak is linked to the stroke volume. The dicrotic notch and diastolic peak are caused by wave reflections, with their location and timing influenced by arterial stiffness [382]. The diastolic decay is determined by the exponential contraction of the arterial system due to outflow of blood. Its time constant, τ, is determined by the systemic vascular resistance, R, and the arterial compliance, C.

Appendix: Additional Results

Some of the results in chapters 5 and 6 were summarised for ease of reading, rather than being given in full. Additional results tables are now provided for completeness.

A.1 Results Relating to Chapter 5

Several comparisons were performed to assess the impact of technical and physiological factors on the quality of respiratory signals extracted from the ECG and PPG in this chapter. The following tables provide additional results for some of the comparisons:

- **PPG measurement site:** Table A.1 provides the results of comparisons between respiratory signals extracted from finger and ear PPG signals (referenced in Section 5.4.1).
- **Signal acquisition equipment:** Table A.2 provides the results of comparisons between respiratory signals exttracted from signals acquired using laboratory and clinical equipment (referenced in Section 5.4.2).
- **Feature- and filter-based techniques:** Table A.3 provides the results of comparisons between feature- and filter-based techniques for extraction of the three fundamental respiratory modulations (referenced in Section 5.4.3).
- **Age:** Tables A.4 and A.5 provide the results of comparisons between young and elderly subjects performed on the RRest-healthy and Fantasia Datasets respectively (referenced in Section 5.4.5).

TABLE A.1: Comparison of qualities of respiratory signals acquired at finger and ear PPG measurement sites: Subject-specific differences in correlation coefficients (CCs) of respiratory signals extracted from finger (fin) and ear PPG signals are expressed as median (lower - upper quartiles) Fin - Ear PPG CCs. Respiratory signals with significantly greater CCs are indicated by a *.

Respiratory signal	Subject-specific differences in CCs	
X_{A1} (BW)	0.06 (-0.01 - 0.17)	***Fin**
X_{A2} (AM)	-0.01 (-0.09 - 0.05)	
X_{A3} (FM)	0.00 (-0.01 - 0.01)	
X_{A4} (BW, AM, FM)	0.01 (-0.02 - 0.04)	
X_{B1} (BW)	0.07 (0.01 - 0.19)	***Fin**
X_{B2} (AM)	0.03 (-0.03 - 0.14)	
X_{B3} (FM)	0.11 (0.04 - 0.18)	***Fin**
X_{B4} (BW)	0.06 (0.00 - 0.14)	***Fin**
X_{B5} (BW, AM)	0.08 (-0.01 - 0.16)	***Fin**
X_{B6} (BW, AM)	0.08 (0.00 - 0.18)	***Fin**
X_{B9} (FM)	0.02 (-0.03 - 0.09)	

TABLE A.2: Comparison of qualities of respiratory signals acquired using laboratory and clinical equipment: Subject-specific differences in correlation coefficients (CCs) of respiratory signals extracted from clinical (clin) and laboratory (lab) equipment are expressed as median (lower - upper quartiles). Respiratory signals with significantly greater CCs are indicated by a *.

Respiratory signal	Subject-specific differences in CCs			
	ECG		PPG	
X_{A1} (BW)	0.01 (-0.05 - 0.10)		0.05 (-0.02 - 0.15)	***Clin**
X_{A2} (AM)	0.01 (-0.01 - 0.05)	***Clin**	-0.02 (-0.07 - 0.03)	
X_{A3} (FM)	0.00 (0.00 - 0.01)		-0.01 (-0.04 - 0.00)	***Lab**
X_{A4} (BW, AM, FM)	0.02 (-0.05 - 0.04)		0.07 (0.03 - 0.11)	***Clin**
X_{B1} (BW)	0.04 (-0.01 - 0.13)	***Clin**	0.07 (-0.01 - 0.15)	***Clin**
X_{B2} (AM)	0.01 (-0.02 - 0.10)		-0.09 (-0.16 - 0.01)	***Lab**
X_{B3} (FM)	0.00 (0.00 - 0.00)		0.02 (-0.01 - 0.07)	***Clin**
X_{B4} (BW)	0.02 (-0.03 - 0.11)		0.02 (-0.05 - 0.10)	
X_{B5} (BW,AM)	0.04 (0.00 - 0.11)	***Clin**	0.01 (-0.08 - 0.12)	
X_{B6} (BW,AM)	0.03 (-0.04 - 0.10)		0.14 (0.04 - 0.20)	***Clin**
X_{B7} (FM)	0.00 (-0.06 - 0.05)		NA	
X_{B8} (AM,FM)	0.01 (-0.06 - 0.08)		NA	
X_{B9} (FM)	0.01 (-0.03 - 0.09)		0.04 (-0.06 - 0.13)	

TABLE A.3: Comparison of feature- and filter-based techniques for extraction of respiratory signals: Subject-specific differences in correlation coefficients (CCs) of respiratory signals extracted using filter-based and feature-based techniques are expressed as median (lower - upper quartiles) feature - filter-based CCs. Respiratory signals with significantly greater CCs are indicated by a *.

Respiratory signals	Subject-specific differences in CCs	
ECG(X_{B1}) - ECG(X_{A1}) (BW)	0.04 (-0.05 - 0.17)	***Feature-based**
ECG(X_{B2}) - ECG(X_{A2}) (AM)	0.34 (0.15 - 0.50)	***Feature-based**
ECG(X_{B3}) - ECG(X_{A3}) (FM)	0.16 (0.06 - 0.21)	***Feature-based**
PPG(X_{B1}) - PPG(X_{A1}) (BW)	0.00 (-0.03 - 0.03)	
PPG(X_{B2}) - PPG(X_{A2}) (AM)	0.06 (-0.01 - 0.11)	***Feature-based**
PPG(X_{B3}) - PPG(X_{A3}) (FM)	0.12 (0.05 - 0.18)	***Feature-based**

TABLE A.4: Comparison of qualities of respiratory signals obtained from young and elderly subjects in the RRest-healthy dataset: Subject-specific correlation coefficients (CCs) for young and elderly subjects, expressed as median (lower - upper quartiles). Respiratory signals with significantly different CCs are indicated by a *.

Respiratory signal	Subject-specific CCs		
	Young	Elderly	
ECG(X_{A1})	0.52 (0.44 - 0.76)	0.60 (0.42 - 0.80)	
ECG(X_{A2})	0.32 (0.23 - 0.58)	0.37 (0.30 - 0.53)	
ECG(X_{A3})	0.43 (0.34 - 0.60)	0.31 (0.25 - 0.47)	
ECG(X_{A4})	0.39 (0.31 - 0.45)	0.37 (0.32 - 0.41)	
ECG(X_{B1})	0.66 (0.57 - 0.77)	0.72 (0.64 - 0.79)	
ECG(X_{B2})	0.76 (0.68 - 0.82)	0.77 (0.68 - 0.81)	
ECG(X_{B3})	0.66 (0.52 - 0.75)	0.44 (0.35 - 0.63)	
ECG(X_{B4})	0.52 (0.42 - 0.68)	0.56 (0.43 - 0.82)	
ECG(X_{B5})	0.74 (0.64 - 0.79)	0.76 (0.69 - 0.78)	
ECG(X_{B6})	0.59 (0.48 - 0.69)	0.50 (0.41 - 0.80)	
ECG(X_{B7})	0.42 (0.36 - 0.51)	0.41 (0.30 - 0.44)	
ECG(X_{B8})	0.73 (0.61 - 0.79)	0.66 (0.44 - 0.76)	
ECG(X_{B9})	0.40 (0.34 - 0.56)	0.36 (0.31 - 0.45)	
PPG(X_{A1})	0.44 (0.38 - 0.57)	0.51 (0.34 - 0.65)	
PPG(X_{A2})	0.41 (0.29 - 0.57)	0.37 (0.27 - 0.53)	
PPG(X_{A3})	0.44 (0.33 - 0.61)	0.30 (0.25 - 0.40)	
PPG(X_{A4})	0.37 (0.30 - 0.42)	0.38 (0.29 - 0.41)	
PPG(X_{B1})	0.48 (0.35 - 0.56)	0.49 (0.36 - 0.63)	
PPG(X_{B2})	0.48 (0.39 - 0.57)	0.47 (0.35 - 0.57)	
PPG(X_{B3})	0.64 (0.50 - 0.73)	0.38 (0.32 - 0.55)	***Young >Elderly**
PPG(X_{B4})	0.45 (0.30 - 0.54)	0.48 (0.28 - 0.55)	
PPG(X_{B5})	0.46 (0.35 - 0.54)	0.51 (0.34 - 0.58)	
PPG(X_{B6})	0.55 (0.44 - 0.61)	0.54 (0.41 - 0.63)	
PPG(X_{B9})	0.41 (0.33 - 0.48)	0.32 (0.28 - 0.52)	

TABLE A.5: Comparison of qualities of respiratory signals obtained from young and elderly subjects in the Fantasia dataset: Subject-specific correlation coefficients (CCs) for young and elderly subjects, expressed as median (lower - upper quartiles). There were no significant differences in the quality of respiratory signals between the two groups.

Respiratory signal	Subject-specific CCs	
	Young	Elderly
ECG(X_{A1})	0.54 (0.46 - 0.64)	0.58 (0.45 - 0.66)
ECG(X_{B1})	0.55 (0.48 - 0.67)	0.55 (0.44 - 0.66)
ECG(X_{B2})	0.68 (0.47 - 0.73)	0.67 (0.43 - 0.78)
ECG(X_{B3})	0.72 (0.65 - 0.79)	0.59 (0.42 - 0.74)
ECG(X_{B4})	0.52 (0.41 - 0.57)	0.51 (0.42 - 0.59)
ECG(X_{B5})	0.56 (0.50 - 0.74)	0.69 (0.43 - 0.74)
ECG(X_{B6})	0.55 (0.45 - 0.65)	0.46 (0.38 - 0.59)
ECG(X_{B7})	0.35 (0.33 - 0.40)	0.34 (0.31 - 0.40)
ECG(X_{B8})	0.49 (0.41 - 0.70)	0.53 (0.44 - 0.68)

A.2 Results Relating to Chapter 6

The performances of RR algorithms were assessed in healthy subjects and hospitalised patients in this chapter. The following additional results tables are provided:

- **Comparison of ECG and PPG input signals:** Tables A.6 and A.7 show how the performances of RR algorithms varied with the use of ECG or PPG as the input signal in healthy and hospitalised subjects respectively (referenced in Sections 6.3.4 and 6.3.5 respectively).
- **Comparison between age groups:** Table A.8 shows how RR algorithms' performances varied between young and elderly healthy subjects (referenced in Section 6.3.4).

TABLE A.6: Comparison of RR algorithm performances when using the ECG or PPG as the input signal in health. Significant differences are annotated by *, with the input signal providing improved performance labelled.

		Algorithm performance, med (quartiles)		
Dataset	Statistic	ECG	PPG	
RRest-healthy	C	4.1 (1.8 - 6.3)	1.0 (0.7 - 1.3)	***ECG**
	$2SD$	7.9 (5.9 - 8.8)	10.5 (9.5 - 11.9)	***ECG**
	bias	-1.2 (-1.9 - -0.4)	-4.5 (-5.5 - -2.3)	***ECG**
	CP_2	61.1 (47.8 - 67.3)	36.5 (33.0 - 41.7)	***ECG**
	iCP_5	15.9 (10.4 - 26.0)	38.7 (30.7 - 48.2)	***ECG**
	TDI_{95}	9.6 (7.0 - 11.9)	15.7 (11.6 - 16.5)	***ECG**
	prop	100.0 (97.5 - 100.0)	100.0 (89.5 - 100.0)	***PPG**

TABLE A.7: Comparison of algorithm performances when using the ECG or PPG as the input signal in hospitalised patients. Significant differences are annotated by *, with the input signal providing improved performance labelled.

		Algorithm performance, med (quartiles)		
Dataset	Statistic	ECG	PPG	
CapnoBase	C	4.4 (2.3 - 11.5)	4.4 (1.7 - 6.4)	
	$2SD$	9.6 (7.9 - 11.3)	9.9 (7.4 - 11.5)	
	bias	-0.1 (-1.0 - 1.3)	1.0 (0.2 - 2.7)	
	CP_2	65.2 (50.1 - 84.5)	73.7 (46.8 - 80.6)	
	iCP_5	14.9 (7.5 - 23.5)	17.2 (12.3 - 28.7)	
	TDI_{95}	10.3 (7.6 - 13.6)	12.5 (9.7 - 15.8)	***ECG**
	prop	98.7 (94.7 - 98.7)	98.7 (88.1 - 98.7)	***PPG**
MIMIC-II	C	1.7 (1.1 - 2.5)	1.3 (0.8 - 2.0)	***ECG**
	$2SD$	11.5 (8.9 - 13.7)	12.4 (9.9 - 13.8)	
	bias	-1.2 (-2.4 - 0.9)	-3.0 (-4.5 - -1.7)	***ECG**
	CP_2	46.8 (37.0 - 57.1)	47.8 (33.7 - 52.0)	
	iCP_5	29.2 (21.5 - 37.5)	38.3 (24.2 - 46.1)	***ECG**
	TDI_{95}	14.9 (11.3 - 17.3)	16.8 (11.1 - 18.7)	
	prop	100.0 (97.6 - 100.0)	98.9 (72.6 - 100.0)	

TABLE A.8: Comparison of RR algorithm performances between young and elderly subjects in health. Significant differences are annotated by *, with the cohort providing improved performance labelled.

Signal & Dataset	Statistic	Algorithm performance, med (quartiles) Young	Elderly	
ECG RRest-healthy	C	4.5 (1.9 - 7.5)	3.3 (2.0 - 4.4)	***young**
	$2SD$	7.7 (5.6 - 8.8)	7.7 (6.5 - 9.6)	***young**
	$bias$	-0.9 (-1.8 - -0.0)	-2.0 (-2.7 - -1.2)	***young**
	CP_2	63.4 (50.5 - 69.9)	56.2 (46.3 - 61.4)	***young**
	iCP_5	14.4 (9.1 - 26.1)	16.8 (13.6 - 25.1)	***young**
	TDI_{95}	9.4 (6.6 - 11.5)	9.9 (7.7 - 11.9)	***young**
	$prop$	100.0 (97.6 - 100.0)	100.0 (96.8 - 100.0)	***elderly**
ECG Fantasia	C	2.2 (1.3 - 3.5)	2.9 (1.4 - 5.5)	***elderly**
	$2SD$	7.8 (6.8 - 10.1)	7.2 (6.2 - 9.1)	***elderly**
	$bias$	-2.4 (-3.8 - -1.2)	-1.9 (-3.1 - -1.0)	***elderly**
	CP_2	50.5 (44.8 - 61.4)	54.8 (46.4 - 69.5)	***elderly**
	iCP_5	22.3 (16.1 - 37.5)	19.2 (12.4 - 35.2)	***elderly**
	TDI_{95}	10.6 (8.0 - 13.0)	9.8 (7.7 - 11.7)	***elderly**
	$prop$	100.0 (98.1 - 100.0)	100.0 (98.1 - 100.0)	***young**
PPG RRest-healthy	C	1.0 (0.7 - 1.4)	0.85 (0.7 - 1.1)	***young**
	$2SD$	10.5 (9.3 - 11.8)	10.8 (9.7 - 12.1)	***young**
	$bias$	-4.0 (-5.3 - -2.1)	-5.1 (-6.0 - -2.7)	***young**
	CP_2	36.7 (32.6 - 43.5)	34.9 (30.0 - 38.7)	***young**
	iCP_5	36.7 (30.4 - 49.2)	42.3 (32.8 - 48.2)	***young**
	TDI_{95}	15.4 (11.3 - 16.4)	15.8 (11.8 - 16.5)	***young**
	$prop$	100.0 (89.9 - 100.0)	100.0 (89.1 - 100.0)	***young**

Bibliography

[1] A. J. Shinton *et al.*, "Impact of graft thickness on visual acuity after Descemet's stripping endothelial keratoplasty." *The British Journal of Ophthalmology*, vol. 96, no. 2, pp. 246–9, 2012.

[2] P. Charlton *et al.*, "Achieving Clinical Quality from Wireless Sensors," in *Conf Proc Eng Med Biol Soc.* Chicago, IL: IEEE, 2014.

[3] T. Bonnici *et al.*, "Continuous Physiological Monitoring of Ambulatory Patients," in *MEC Annual Meeting and Bioengineering14 Programme and Abstracts.* London: MECbioeng14, Imperial College London, 2014, p. 38.

[4] M. A. F. Pimentel, P. H. Charlton, and D. A. Clifton, "Probabilistic estimation of respiratory rate from wearable sensors," in *Wearable Electronics Sensors*, S. C. Mukhopadhyay, Ed. Springer International Publishing, 2015, vol. 15, pp. 241–62.

[5] P. H. Charlton, "The processes and benefits of sharing clinical data," in *Data Dialogue: Time to Share: Navigating Boundaries*, Cambridge, UK, 2016.

[6] P. H. Charlton, M. Villarroel, and F. Salguiero, "Waveform analysis to estimate respiratory rate," in *Secondary Analysis of Electronic Health Records.* Springer International Publishing, 2016, ch. 26, pp. 377–90.

[7] P. H. Charlton *et al.*, "Extraction of respiratory signals from the electrocardiogram and photoplethysmogram: technical and physiological determinants," *Physiological Measurement*, vol. 38, no. 5, pp. 669–90, 2017.

[8] P. Charlton *et al.*, "The influence of recording equipment on the accuracy of respiratory rate estimation from the electrocardiogram and photoplethysmogram," in *MEC Annual Meeting and Bioengineering14 Programme and Abstracts.* London: MECbioeng14, Imperial College London, 2014, p. 96.

[9] P. Charlton *et al.*, "An Assessment of Algorithms to Estimate Respiratory Rate from the Electrocardiogram and Photoplethysmogram," in *MEIBioeng16*, Oxford, 2016.

[10] P. H. Charlton *et al.*, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiological Measurement*, vol. 37, no. 4, pp. 610–26, 2016.

[11] P. H. Charlton, M. Pimentel, and S. Lokhandwala, "Data Fusion Techniques for Early Warning of Clinical Deterioration," in *Secondary Analysis of Electronic Health Records*. Springer International Publishing, 2016, pp. 325–38.

[12] P. McAleese and W. Odling-Smee, "The effect of complications on length of stay," *Annals of Surgery*, vol. 220, no. 6, pp. 740–4, 1994.

[13] N. A. Khan *et al.*, "Association of postoperative complications with hospital costs and length of stay in a tertiary care center," *Journal of General Internal Medicine*, vol. 21, no. 2, pp. 177–80, 2006.

[14] R. J. Lagoe, P. E. Johnson, and M. P. Murphy, "Inpatient hospital complications and lengths of stay: a short report," *BMC Research Notes*, vol. 4, no. 1, p. 135, 2011.

[15] E. J. Thomas *et al.*, "Costs of medical injuries in Utah and Colorado," *Inquiry*, vol. 36, no. 3, pp. 255–64, 1999.

[16] R. M. Wilson *et al.*, "The Quality in Australian Health Care Study," *The Medical Journal of Australia*, vol. 163, no. 9, pp. 458–71, 1995.

[17] U. Kyriacos, J. Jelsma, and S. Jordan, "Monitoring vital signs using early warning scoring systems: a review of the literature." *Journal of Nursing Management*, vol. 19, no. 3, pp. 311–30, 2011.

[18] R. M. Schein *et al.*, "Clinical antecedents to in-hospital cardiopulmonary arrest," *Chest*, vol. 98, no. 6, pp. 1388–92, 1990.

[19] C. Franklin and J. Mathew, "Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event," *Critical Care Medicine*, vol. 22, no. 2, pp. 244–7, 1994.

[20] M. D. Buist *et al.*, "Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care. A pilot study in a tertiary-care hospital," *The Medical Journal of Australia*, vol. 171, no. 1, pp. 22–5, 1999.

[21] K. M. Hillman *et al.*, "Duration of life-threatening antecedents prior to intensive care admission," *Intensive Care Medicine*, vol. 28, no. 11, pp. 1629–34, 2002.

[22] K. M. Hillman *et al.*, "Antecedents to hospital deaths," *Internal Medicine Journal*, vol. 31, no. 6, pp. 343–8, 2001.

[23] C. A. Alvarez *et al.*, "Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data," *BMC Medical Informatics and Decision Making*, vol. 13, no. 28, 2013.

[24] M. a. DeVita *et al.*, ""Identifying the hospitalised patient in crisis"–a consensus conference on the afferent limb of rapid response systems." *Resuscitation*, vol. 81, no. 4, pp. 375–82, 2010.

[25] A. F. Smith and J. Wood, "Can some in-hospital cardio-respiratory arrests be prevented? A prospective survey," *Resuscitation*, vol. 37, no. 3, pp. 133–7, 1998.

[26] Patient Safety Observatory, "Safer care for the acutely ill patient: learning from serious incidents," National Patient Safety Agency, London, Tech. Rep., 2007.

[27] J. Whittington *et al.*, "Using an automated risk assessment report to identify patients at risk for clinical deterioration," *Joint Commission Journal on Quality and Patient Safety*, vol. 33, no. 9, pp. 569–74, 2007.

[28] Royal College of Physicians, "National Early Warning Score (NEWS): Standardising the assessment of acute-illness severity in the NHS," Report of a working party. London: RCP, 2012.

[29] V. Nangalia, D. R. Prytherch, and G. B. Smith, "Health technology assessment review: remote monitoring of vital signs – current status and future challenges," *Critical Care*, vol. 14, no. 5, p. 233, 2010.

[30] R. J. Cusack and J. F. Coutts, "Do not be alarmed, the patient is monitored," *Critical Care*, vol. 7, no. 5, pp. 349–50, 2003.

[31] C. P. Subbe *et al.*, "Effect of introducing the Modified Early Warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions." *Anaesthesia*, vol. 58, no. 8, pp. 797–802, 2003.

[32] P. B. Lovett *et al.*, "The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage," *Annals of Emergency Medicine*, vol. 45, no. 1, pp. 68–76, 2005.

[33] J. Ko *et al.*, "Wireless sensing systems in clinical environments: improving the efficiency of the patient monitoring process," *IEEE Engineering in Medicine and Biology Magazine*, vol. 29, no. 2, pp. 103–9, 2010.

[34] D.-D. Cale, "A new perspective on patient monitoring," *Nursing Management*, vol. 38, no. 12, pp. 24–6, 2007.

[35] P. H. Charlton, "Monitoring physiological trajectories," Transfer of status report, King's College London, 2014.

[36] D. Jones *et al.*, "Defining clinical deterioration," *Resuscitation*, vol. 84, no. 8, pp. 1029–34, 2013.

[37] P. J. Watkinson *et al.*, "A randomised controlled trial of the effect of continuous electronic physiological monitoring on the adverse event rate in high risk medical and surgical patients," *Anaesthesia*, vol. 61, no. 11, pp. 1031–9, 2006.

[38] R. Bellomo *et al.*, "Postoperative serious adverse events in a teaching hospital: a prospective study," *The Medical Journal of Australia*, vol. 176, no. 5, pp. 216–8, 2002.

[39] M. A. Healey *et al.*, "Complications in surgical patients," *Archives of Surgery*, vol. 137, no. 5, pp. 611–7; discussion 617–8, 2002.

[40] J. H. Silber *et al.*, "Evaluation of the complication rate as a measure of quality of care in coronary artery bypass graft surgery," *Journal of the American Medical Association*, vol. 274, no. 4, pp. 317–23, 1995.

[41] D. Dindo, N. Demartines, and P.-A. Clavien, "Classification of Surgical Complications," *Annals of Surgery*, vol. 240, no. 2, pp. 205–13, 2004.

[42] K. Hillman *et al.*, "Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial," *Lancet*, vol. 365, no. 9477, pp. 2091–7, 2005.

[43] J. Chen *et al.*, "The relationship between early emergency team calls and serious adverse events." *Critical Care Medicine*, vol. 37, no. 1, pp. 148–53, 2009.

[44] A. J. Bleyer *et al.*, "Longitudinal analysis of one million vital signs in patients in an academic medical center," *Resuscitation*, vol. 82, no. 11, pp. 1387–92, 2011.

[45] G. B. Smith *et al.*, "The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death," *Resuscitation*, vol. 84, no. 4, pp. 465–70, 2013.

[46] B. H. Cuthbertson and G. B. Smith, "A warning on early-warning scores!" *British Journal of Anaesthesia*, vol. 98, no. 6, pp. 704–6, 2007.

[47] D. R. Goldhill *et al.*, "A physiologically-based early warning score for ward patients: the association between score and outcome," *Anaesthesia*, vol. 60, no. 6, pp. 547–53, 2005.

[48] R. Paterson *et al.*, "Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit," *Clinical Medicine*, vol. 6, no. 3, pp. 281–4, 2006.

[49] M. M. Churpek *et al.*, "Predicting cardiac arrest on the wards: a nested case-control study," *Chest*, vol. 141, no. 5, pp. 1170–6, 2012.

[50] R. Bellomo *et al.*, "A controlled trial of electronic automated advisory vital signs monitoring in general hospital wards," *Critical Care Medicine*, vol. 40, no. 8, pp. 2349–61, 2012.

[51] M. Alvi, "Modelling Physiological Patterns in Post-operative Cancer Patients for Early Warning of Deterioration," MEng. Dissertation, Unviersity of Oxford, 2013.

[52] H. Gao *et al.*, "Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward." *Intensive Care Medicine*, vol. 33, no. 4, pp. 667–79, 2007.

[53] G. B. Smith *et al.*, "Review and performance evaluation of aggregate weighted 'track and trigger' systems." *Resuscitation*, vol. 77, no. 2, pp. 170–9, 2008.

[54] M. P. Turakhia *et al.*, "Latency of ECG displays of hospital telemetry systems: a science advisory from the American Heart Association." *Circulation*, vol. 126, no. 13, pp. 1665–9, 2012.

[55] B. J. Drew *et al.*, "Practice Standards for Electrocardiographic Monitoring in Hospital Settings: an American Heart Association Scientific Statement from the Councils on Cardiovascular Nursing, Clinical Cardiology, and Cardiovascular Disease in the Young," *Circulation*, vol. 110, no. 17, pp. 2721–46, 2004.

[56] M. J. Schull and D. A. Redelmeier, "Continuous electrocardiographic monitoring and cardiac arrest outcomes in 8,932 telemetry ward patients," *Academic Emergency Medicine*, vol. 7, no. 6, pp. 647–652, 2000.

[57] W. J. Brady *et al.*, "In-hospital cardiac arrest: impact of monitoring and witnessed event on patient survival and neurologic status at hospital discharge." *Resuscitation*, vol. 82, no. 7, pp. 845–52, 2011.

[58] K. Cleverley *et al.*, "The impact of telemetry on survival of in-hospital cardiac arrests in non-critical care patients." *Resuscitation*, vol. 84, no. 7, pp. 878–82, 2013.

[59] P. S. Chan *et al.*, "Delayed time to defibrillation after in-hospital cardiac arrest." *The New England Journal of Medicine*, vol. 358, no. 1, pp. 9–17, 2008.

[60] N. Najafi and A. Auerbach, "Use and outcomes of telemetry monitoring on a medicine service," *Archives of Internal Medicine*, vol. 172, no. 17, pp. 1349–1350, 2012.

[61] C. A. Estrada *et al.*, "Role of telemetry monitoring in the non-intensive care unit." *The American Journal of Cardiology*, vol. 76, no. 12, pp. 960–5, 1995.

[62] J. E. Hollander *et al.*, "Are monitored telemetry beds necessary for patients with nontraumatic chest pain and normal or nonspecific electrocardiograms?" *The American Journal of Cardiology*, vol. 79, no. 8, pp. 1110–1, 1997.

[63] S. A. Grossman *et al.*, "Is telemetry useful in evaluating chest pain patients in an observation unit?" *Internal and Emergency Medicine*, vol. 6, no. 6, pp. 543–6, 2011.

[64] C. Atzema *et al.*, "ALARMED: adverse events in low-risk patients with chest pain receiving continuous electrocardiographic monitoring in the emergency department. A pilot study." *The American Journal of Emergency Medicine*, vol. 24, no. 1, pp. 62–7, 2006.

[65] E. H. Chen, "Appropriate Use of Telemetry Monitoring in Hospitalized Patients," *Current Emergency and Hospital Medicine Reports*, vol. 2, no. 1, pp. 52–6, 2013.

[66] C. Orphanidou *et al.*, "Telemetry-based vital sign monitoring for ambulatory hospital patients," in *Conf Proc Eng Med Biol Soc.* IEEE, 2009, pp. 4650–3.

[67] M. Hernandez-Silveira *et al.*, "Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs," *BMJ Open*, vol. 5, no. 5, p. e006606, 2015.

[68] M. Hernandez-Silveira *et al.*, "Preliminary assessment of the SensiumVitals®: A low-cost wireless solution for patient surveillance in the general wards," in *Conf Proc Eng Med Biol Soc.* IEEE, 2015, pp. 4931–7.

[69] J. Welch, J. Moon, and S. McCombie, "Early detection of the deteriorating patient: the case for a multi-parameter patient-worn monitor," *Biomedical Instrumentation & Technology*, vol. 46, no. s2, pp. 57–64, 2012.

[70] L. Tarassenko, A. Hann, and A. Patterson, "Biosign: Multi-parameter monitoring for early warning of patient deterioration," in *Proc Int Sem Medical Applications of Signal Processing.* IEE, 2005, pp. 71–6.

[71] F. Ismail and M. Davies, "Integrated monitoring and analysis for early warning of patient deterioration." *British Journal of Anaesthesia*, vol. 98, no. 1, pp. 149–50; author reply 150–2, 2007.

[72] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–6, 1985.

[73] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Transactions on Biomedical Engineering*, vol. 33, no. 12, pp. 1157–65, 1986.

[74] W. Karlen, J. M. Ansermino, and G. Dumont, "Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications," in *Conf Proc Eng Med Biol Soc.* IEEE, 2012, pp. 3131–4.

[75] T. Bonnici *et al.*, "Testing of Wearable Monitors in a Real-World Hospital Environment: What Lessons Can Be Learnt?" in *Conf Proc 9th Wearable and Implantable BSNs.* IEEE, 2012, pp. 79–84.

[76] A. Garbino *et al.*, "Physiological monitoring and analysis of a manned stratospheric balloon test program," *Aviation, Space, and Environmental Medicine*, vol. 85, no. 2, pp. 177–82, 2014.

[77] P. Hubner *et al.*, "Surveillance of patients in the waiting area of the department of emergency medicine," *Medicine*, vol. 94, no. 51, p. e2322, 2015.

[78] J. Hernandez, D. McDuff, and R. Picard, "BioWatch: Estimation of Heart and Breathing Rates from Wrist Motions," in *Conf Proc Pervasive Computing Technologies for Healthcare.*, 2015, pp. 169–76.

[79] R. Bailon, L. Sornmo, and P. Laguna, "ECG-derived respiratory frequency estimation," in *Advanced methods and tools for ECG data analysis*, G. D. Clifford, F. Azuaje, and P. E. McSharry, Eds. London: Artech House, 2006, ch. 8, pp. 215–44.

[80] W. Karlen *et al.*, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1946–53, 2013.

[81] C. Toumazou and P. Georgiou, "Bio-inspired semiconductors for early detection and therapy," in *Conf Proc Asian Solid-State Circuits.* IEEE, 2011, pp. 129–32.

[82] N. Fålun *et al.*, "Evaluation of the appropriateness and outcome of in-hospital telemetry monitoring." *The American Journal of Cardiology*, vol. 112, no. 8, pp. 1219–23, 2013.

[83] Koninklijke Philips N.V., "Philips IntelliVue MX40 Patient Monitor Brochure: Monitoring on the go," 2016.

[84] D. W. Curtis *et al.*, "SMART–an integrated wireless system for monitoring unattended patients," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 44–53, 2008.

[85] D. Kisner *et al.*, "Reduced incidence of atrial fibrillation after cardiac surgery by continuous wireless monitoring of oxygen saturation on the normal ward and resultant oxygen therapy for hypoxia," *European Journal of Cardiothoracic Surgery*, vol. 35, no. 1, pp. 111–5, 2009.

[86] A. O. Chipara *et al.*, "Reliable Patient Monitoring : A Clinical Study in a Step-down Hospital Unit," *Technical Report, Dept. Comput. Sci. Eng., Washington Univ. St. Louis*, 2009.

[87] O. Chipara *et al.*, "Reliable Clinical Monitoring using Wireless Sensor Networks : Experiences in a Step-down Hospital Unit," in *Conf Proc Embedded Networked Sensor Systems.* ACM, 2010.

[88] C. V. Pollack, "Wireless cardiac event alert monitoring is feasible and effective in the emergency department and adjacent waiting areas," *Critical Pathways in Cardiology*, vol. 8, no. 1, pp. 7–11, 2009.

[89] J. Ko, J. Lim, and Y. Chen, "MEDiSN: Medical emergency detection in sensor networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 10, no. 1, 2010.

[90] G. López, V. Custodio, and J. I. Moreno, "LOBIN: E-textile and wireless-sensor-network-based platform for healthcare monitoring in future hospital environments." *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 6, pp. 1446–58, 2010.

[91] N. Donnelly *et al.*, "Development of a ubiquitous clinical monitoring solution to improve patient safety and outcomes," in *Conf Proc Eng Med Biol Soc.* IEEE, 2012, pp. 6068–73.

[92] R. Dor *et al.*, "Experiences with an End-To-End Wireless Clinical Monitoring System," in *Conf Proc Wireless Health.* ACM, 2012.

[93] M. Cretikos *et al.*, "The objective medical emergency team activation criteria: A case-control study," *Resuscitation*, vol. 73, no. 1, pp. 62–72, 2007.

[94] S. E. Bedell *et al.*, "Incidence and characteristics of preventable iatrogenic cardiac arrests," *Journal of the American Medical Association*, vol. 265, no. 21, pp. 2815–20, 1991.

[95] J. F. Fieselmann *et al.*, "Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients," *Journal of General Internal Medicine*, vol. 8, no. 7, pp. 354–60, 1993.

[96] T. J. Hodgetts *et al.*, "The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team." *Resuscitation*, vol. 54, no. 2, pp. 125–31, 2002.

[97] T. R. Gravelyn and J. G. Weg, "Respiratory rate as an indicator of acute respiratory dysfunction." *Journal of the American Medical Association*, vol. 244, no. 10, pp. 1123–5, 1980.

[98] R. W. Duckitt *et al.*, "Worthing physiological scoring system: derivation and validation of a physiological early-warning system for medical admissions. An observational, population-based single-centre study." *British Journal of Anaesthesia*, vol. 98, no. 6, pp. 769–74, 2007.

[99] M. Buist *et al.*, "Association between clinically abnormal observations and subsequent in-hospital mortality: a prospective study." *Resuscitation*, vol. 62, no. 2, pp. 137–41, 2004.

[100] World Health Organization (WHO), "Fourth Programme Report, 1988-1989: ARI Programme for Control of Acute Respiratory Infections," WHO, Geneva, Tech. Rep., 1990.

[101] A. Chellel *et al.*, "Nursing observations on ward patients at risk of critical illness," *Nursing Times*, vol. 98, no. 46, pp. 36–9, 2002.

[102] J. Hogan, "Why don't nurses monitor the respiratory rates of patients?" *British Journal of Nursing*, vol. 15, no. 9, pp. 489–92, 2006.

[103] K. Philip, R. Richardson, and M. Cohen, "Staff perceptions of respiratory rate measurement in a general hospital," *British Journal of Nursing*, vol. 22, no. 10, pp. 570–4, 2007.

[104] M. A. Cretikos *et al.*, "Respiratory rate: the neglected vital sign," *The Medical Journal of Australia*, vol. 188, no. 11, pp. 657–9, 2008.

[105] M. Młyńczak and G. Cybulski, "Impedance pneumography: is it possible?" in *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2012*, R. S. Romaniuk, Ed., vol. 8454, 2012, p. 84541T.

[106] B. G. Goudra, "Comparison of Acoustic Respiration Rate, Impedance Pneumography and Capnometry Monitors for Respiration Rate Accuracy and Apnea Detection during GI Endoscopy Anesthesia," *Open Journal of Anesthesiology*, vol. 3, pp. 74–9, 2013.

[107] S. Lapi *et al.*, "Respiratory rate assessments using a dual-accelerometer device," *Respiratory Physiology and Neurobiology*, vol. 191, no. 1, pp. 60–6, 2014.

[108] S. G. Fleming and L. Tarassenko, "A comparison of signal processing techniques for the extraction of breathing rate from the photoplethysmogram," *International Journal of Biological and Life Sciences*, vol. 2, no. 4, pp. 233–7, 2006.

[109] A. Schäfer and K. W. Kratky, "Estimation of breathing rate from respiratory sinus arrhythmia: comparison of various methods," *Annals of Biomedical Engineering*, vol. 36, no. 3, pp. 476–85, 2008.

[110] A. Garde *et al.*, "Estimating respiratory and heart rates from the correntropy spectral density of the photoplethysmogram," *PLoS ONE*, vol. 9, no. 1, p. e86427, 2014.

[111] S. Nemati, A. Malhotra, and G. D. Clifford, "Data fusion for improved respiration rate estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 926305, 2010.

[112] R. Ruangsuwana, G. Velikic, and M. Bocko, "Methods to extract respiration information from ECG signals," in *Conf Proc ICASSP.* IEEE, 2010, pp. 570–3.

[113] J. Lázaro *et al.*, "Deriving respiration from photoplethysmographic pulse width," *Medical and Biological Engineering and Computing*, vol. 51, no. 1-2, pp. 233–42, 2013.

[114] S. Nizami, J. R. Green, and C. McGregor, "Implementation of artifact detection in critical care: a methodological review," *IEEE Reviews in Biomedical Engineering*, vol. 6, pp. 127–42, 2013.

[115] L. Tarassenko, A. Hann, and D. Young, "Integrated monitoring and analysis for early warning of patient deterioration," *British Journal of Anaesthesia*, vol. 97, no. 1, pp. 64–8, 2006.

[116] L. a. Lynn and J. P. Curry, "Patterns of unexpected in-hospital deaths: a root cause analysis," *Patient Safety in Surgery*, vol. 5, no. 1, p. 3, 2011.

[117] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomedical Instrumentation & Technology*, vol. 46, no. 4, pp. 268–77, 2012.

[118] B. Gross, D. Dahl, and L. Nielsen, "Physiologic monitoring alarm load on medical/surgical floors of a community hospital." *Biomedical Instrumentation & Technology*, vol. Suppl, pp. 29–36, 2011.

[119] J. P. Keller, "Clinical alarm hazards: A top ten health technology safety concern," *Journal of Electrocardiology*, vol. 45, no. 6, pp. 588–91, 2012.

[120] B. Bridgewater *et al.*, "National Adult Cardiac Surgery Audit Report," Institute for Cardiovascular Outcomes Research, London, Tech. Rep., 2012.

[121] Y. Lee *et al.*, "the Usefulness of Modified National Early Warning Score With the Age Level in Critically Ill Medical Patients," *Intensive Care Medicine Experimental*, vol. 3, no. Suppl 1, p. A834, 2015.

[122] M. M. Churpek *et al.*, "Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards." *Critical Care Medicine*, vol. 42, no. 4, pp. 841–8, 2014.

[123] T. Bonnici *et al.*, "Experiences implementing a system for widespread recording of patient physiology in Clinical Practice, Research and Sepsis poster Presentations," *Journal of the Intensive Care Society*, vol. 16, no. 1_suppl, pp. 85–6, 2015.

[124] C. Orphanidou *et al.*, "Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 832–8, 2015.

[125] J. X. Sun, "Cardiac Output Estimation using Arterial Blood Pressure Waveforms," MEng. Thesis, Massachusetts Institute of Technology, 2006.

[126] S. J. Redmond *et al.*, "Electrocardiogram signal quality measures for unsupervised telehealth environments," *Physiological Measurement*, vol. 33, no. 9, pp. 1517–33, 2012.

[127] G. Clifford, "rpeakdetect.m," Available at: http://www.mit.edu/~gari/CODE/ECGtools.

[128] M. Pimentel, "adaptPulseSegment.m," Available at: https://github.com/peterhcharlton/RRest.

[129] W. Zong *et al.*, "An open-source algorithm to detect onset of arterial blood pressure pulses," in *Proc CinC*. IEEE, 2003, pp. 259–62.

[130] M. M. Churpek *et al.*, "Multicenter Development and Validation of a Risk Stratification Tool for Ward Patients," *American Journal of Respiratory and Critical Care Medicine*, vol. 190, pp. 649–655, 2014.

[131] A. J. Seely *et al.*, "Do heart and respiratory rate variability improve prediction of extubation outcomes in critically ill patients?" *Critical Care*, vol. 18, p. R65, 2014.

[132] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

[133] N. Gilboy *et al.*, "Emergency severity index, version 4: implementation handbook," Rockville, MD, 2005.

[134] N. Farrohknia *et al.*, "Emergency department triage scales and their components: a systematic review of the scientific evidence," *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, vol. 19, p. 42, 2011.

[135] A. Khalil, G. Kelen, and R. E. Rothman, "A simple screening tool for identification of community-acquired pneumonia in an inner city emergency department," *Emergency Medicine Journal*, vol. 24, no. 5, pp. 336–8, 2007.

[136] The ACCP/SCCM Consensus Conference Committee, "American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference: definitions for sepsis and

organ failure and guidelines for the use of innovative therapies in sepsis." *Critical Care Medicine*, vol. 20, no. 6, pp. 864–74, 1992.

[137] S. Z. Goldhaber, L. Visani, and M. De Rosa, "Acute pulmonary embolism: Clinical outcomes in the International Cooperative Pulmonary Embolism Registry (ICOPER)," *Lancet*, vol. 353, no. 2182, pp. 1386–9, 1999.

[138] W. A. Knaus *et al.*, "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults." *Chest*, vol. 100, no. 6, pp. 1619–36, 1991.

[139] L. Tarassenko *et al.*, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiological Measurement*, vol. 35, no. 5, pp. 807–31, 2014.

[140] D. J. Meredith *et al.*, "Photoplethysmographic derivation of respiratory rate: a review of relevant physiology," *Journal of Medical Engineering & Technology*, vol. 36, no. 1, pp. 1–7, 2012.

[141] S. Pikkujamsa, T. Makikallio, and L. Sourander, "Cardiac interbeat interval dynamics from childhood to senescence," *Circulation*, vol. 100, no. 4, pp. 393–9, 1999.

[142] J. M. Moll and V. Wright, "An objective clinical study of chest expansion," *Annals of the Rheumatic Diseases*, vol. 31, no. 1, pp. 1–8, 1972.

[143] I. O'Brien, P. O'Hare, and R. Corrall, "Heart rate variability in healthy subjects: effect of age and the derivation of normal ranges for tests of autonomic function," *British Heart Journal*, vol. 57, no. 1, pp. 109–10, 1986.

[144] M. Nitzan, I. Faib, and H. Friedman, "Respiration-induced changes in tissue blood volume distal to occluded artery, measured by photoplethysmography," *Journal of Biomedical Optics*, vol. 11, no. 4, p. 040506, 2006.

[145] G. G. Berntson, J. T. Cacioppo, and K. S. Quigley, "Respiratory sinus arrhythmia: Autonomic origins, physiological mechanisms, and psychophysiological implications," *Psychophysiology*, vol. 30, no. 2, pp. 183–96, 1993.

[146] P. D. Larsen *et al.*, "Respiratory sinus arrhythmia in conscious humans during spontaneous respiration," *Respiratory Physiology & Neurobiology*, vol. 174, no. 1, pp. 111–8, 2010.

[147] P. S. Addison *et al.*, "Developing an algorithm for pulse oximetry derived respiratory rate (RR(oxi)): a healthy volunteer study," *Journal of Clinical Monitoring and Computing*, vol. 26, no. 1, pp. 45–51, 2012.

[148] P. S. Addison *et al.*, "Pulse oximetry-derived respiratory rate in general care floor patients," *Journal of Clinical Monitoring and Computing*, vol. 29, no. 1, pp. 113–20, 2014.

[149] C. Ahlstrom *et al.*, "A respiration monitor based on electrocardiographic and photoplethysmographic sensor fusion," in *Conf Proc Eng Med Biol Soc.* IEEE, 2004, pp. 2311–4.

[150] S. P. Arunachalam and L. F. Brown, "Real-time estimation of the ECG-derived respiration (EDR) signal using a new algorithm for baseline wander noise removal," in *Conf Proc Eng Med Biol Soc.* IEEE, 2009, pp. 5681–4.

[151] S. Babaeizadeh *et al.*, "A comparison of three ECG-derived respiration techniques during cardiac magnetic resonance image acquisition," *Journal of Electrocardiology*, vol. 45, no. 6, p. 696, 2012.

[152] S. Babaeizadeh and P. Healthcare, "Cost-efficient accurate monitoring of respiration rate using ECG," in *Conf Proc CinC.* IEEE, 2015, pp. 1009–12.

[153] S. Babaeizadeh *et al.*, "Electrocardiogram-derived respiration in screening of sleep-disordered breathing," *Journal of Electrocardiology*, vol. 44, no. 6, pp. 700–6, 2011.

[154] D. Birrenkott *et al.*, "Robust estimation of respiratory rate via ECG- and PPG-derived respiratory quality indices," in *Conf Proc Eng Med Biol Soc.* IEEE, 2016, pp. 676–9.

[155] D. Birrenkott, "Respiratory quality index design and validation for ECG and PPG derived respiratory data," Report for transfer of status, University of Oxford, 2015.

[156] E. Bowers, A. Murray, and P. Langley, "Respiratory rate derived from principal component analysis of single lead electrocardiogram," in *Conf Proc CinC.* IEEE, 2008, pp. 437–40.

[157] J. Boyle *et al.*, "Automatic detection of respiration rate from ambulatory single-lead ECG," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 890–6, 2009.

[158] L. F. Brown and S. P. Arunachalam, "Real-time estimation of the ECG-derived respiration (EDR) signal," in *Biomedical Sciences Instrumentation*, vol. 45, 2009, pp. 59–64.

[159] M. Campolo *et al.*, "ECG-derived respiratory signal using Empirical Mode Decomposition," in *Proc Int Symp Medical Measurements and Applications.* IEEE, 2011, pp. 399–403.

[160] A. Canu and M. Canu, "Embedded low-power system for respiration rate calculation," MSc. Thesis, Unviersity College Cork, 2010.

[161] A. Canu *et al.*, "Respiration rate calculation using low power DSP processor and SpO2 sensor," in *Proc Int Symp Medical Measurements and Applications.* IEEE, 2011, pp. 517–20.

[162] R. A. Cernat *et al.*, "Recording system and data fusion algorithm for enhancing the estimation of the respiratory rate from photoplethysmogram," in *Conf Proc Eng Med Biol Soc.* IEEE, 2015, pp. 5977–80.

[163] A. M. Chan, N. Ferdosi, and R. Narasimhan, "Ambulatory respiratory rate detection using ECG and a triaxial accelerometer," in *Conf Proc Eng Med Biol Soc.* IEEE, 2013, pp. 4058–61.

[164] K. H. Chon, S. Dash, and K. Ju, "Estimation of respiratory rate from photoplethysmogram data using time-frequency spectral estimation," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2054–63, 2009.

[165] S. S. Chreiteh *et al.*, "Estimation of respiratory rates based on photoplethysmographic measurements at the sternum," in *Conf Proc Eng Med Biol Soc.* IEEE, 2015, pp. 6570–3.

[166] D. A. Clifton *et al.*, "Home monitoring: breathing rate from PPG and ECG," Oxford, 2012.

[167] D. Clifton *et al.*, "Measurement of respiratory rate from the photoplethysmogram in chest clinic patients," *Journal of Clinical Monitoring & Computing*, vol. 21, no. 1, pp. 55–61, 2007.

[168] D. Cysarz *et al.*, "Comparison of respiratory rates derived from heart rate variability, ECG amplitude, and nasal/oral airflow," *Annals of Biomedical Engineering*, vol. 36, no. 12, pp. 2085–94, 2008.

[169] S. Dabiri and M. A. M. Shirazi, "Estimation of respiratory rate from photoplethysmogram signal of sleep apnea patients: A comparative study of different methods," in *Conf Proc TSP.* IEEE, 2015, pp. 440–3.

[170] N. Daimiwal, M. Sundhararajan, and R. Shriram, "Respiratory rate, heart rate and continuous measurement of BP using PPG," in *Conf Proc CSP.*, 2014, pp. 999–1002.

[171] G. Dan, Z. Li, and H. Ding, "A mother wavelet selection algorithm for respiratory rate estimation from photoplethysmogram," in *IFMBE Proceedings*, D. A. Jaffray, Ed. Springer International Publishing, 2015, vol. 51, pp. 962–5.

[172] S. Dash *et al.*, "Estimation of respiratory rate from ECG, photoplethysmogram, and piezoelectric pulse transducer signals: a comparative study of time-frequency methods," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1099–107, 2010.

[173] P. Dehkordi *et al.*, "Estimating instantaneous respiratory rate from the photoplethysmogram," in *Conf Proc Eng Med Biol Soc.* IEEE, 2015, pp. 6150–3.

[174] S. Ding *et al.*, "Derivation of respiratory signal from single-channel ECGs based on source statistics," *International Journal of Bioelectromagnetism*, vol. 6, no. 1, 2004.

[175] A. Espiritu Santo and C. Carbajal, "Respiration rate extraction from ECG signal via discrete wavelet transform," in *2nd Circuits and Systems for Medical and Environmental Applications Workshop (CASME).* IEEE, 2010, pp. 1–4.

[176] R. Firoozabadi, E. D. Helfenbein, and S. Babaeizadeh, "Monitoring respiration rate in sleep-disordered breathing patients using chest belts or ECG," *Journal of Electrocardiology*, vol. 47, no. 6, p. 908, 2014.

[177] S. Fleming *et al.*, "Non-invasive measurement of respiratory rate in children using the photoplethysmogram," in *Conf Proc Eng Med Biol Soc.* IEEE, 2008, pp. 1886–9.

[178] J. Y. a. Foo and S. J. Wilson, "Estimation of breathing interval from the photoplethysmographic signals in children," *Physiological Measurement*, vol. 26, no. 6, pp. 1049–58, 2005.

[179] G. D. Furman *et al.*, "Electrocardiogram derived respiration during sleep," in *Conf Proc CinC.* IEEE, 2005, pp. 351–4.

[180] A. Fusco *et al.*, "On how to extract breathing rate from PPG signal using wearable devices," in *Conf Proc BioCAS.* IEEE, 2015, pp. 3–6.

[181] A. Garde *et al.*, "Empirical mode decomposition for respiratory and heart rate estimation from the photoplethysmogram," in *Conf Proc CinC.* Zaragoza: IEEE, 2013, pp. 799–802.

[182] L. Guo, S. Lei, and M. Pan, "The methods and limitations of extracting respiratory rhythm utilizing photoplethysmographic signals," in *Conf Proc BMEI.* IEEE, 2010, pp. 1059–1062.

[183] E. Helfenbein *et al.*, "Electrocardiogram/electromyogram-derived respiration in the presence of simulated respiratory disease," *Journal of Electrocardiology*, vol. 44, no. 6, pp. 751–2, 2011.

[184] E. Helfenbein *et al.*, "ECG/EMG-derived respiration during cardiac magnetic resonance imaging," *Journal of Electrocardiology*, vol. 45, no. 6, pp. 693–4, 2012.

[185] E. Helfenbein *et al.*, "Development of three methods for extracting respiration from the surface ECG: a review," *Journal of Electrocardiology*, vol. 47, no. 6, pp. 819–25, 2014.

[186] A. Johansson, "Neural network for photoplethysmographic respiratory rate monitoring," *Medical & Biological Engineering & Computing*, vol. 41, no. 3, pp. 242–8, 2003.

[187] A. Johansson *et al.*, "Pulse wave transit time for monitoring respiration rate," *Medical & Biological Engineering & Computing*, vol. 44, no. 6, pp. 471–8, 2006.

[188] A. Johansson and P. A. Oberg, "Estimation of respiratory volumes from the photoplethysmographic signal. Part 2: A model study," *Medical & Biological Engineering & Computing*, vol. 37, no. 1, pp. 48–53, 1999.

[189] W. S. Johnston and Y. Mendelson, "Extracting breathing rate information from a wearable reflectance pulse oximeter sensor," in *Conf Proc Eng Med Biol Soc.* IEEE, 2004, pp. 5388–91.

[190] T. Kagawa, A. Kawamoto, and N. Nakajima, "System for simultaneous measurement of breathing rate and heart rate using photoplethysmogram," in *Conf Proc Body Area Networks.* ICST, 2013.

[191] W. Karlen *et al.*, "Respiratory rate estimation using respiratory sinus arrhythmia from photoplethysmography," in *Conf Proc Eng Med Biol Soc.* IEEE, 2011, pp. 1201–4.

[192] W. Karlen *et al.*, "Respiratory rate assessment from photoplethysmographic imaging," in *Conf Proc Eng Med Biol Soc.* IEEE, 2014, pp. 5397–400.

[193] W. Karlen *et al.*, "Estimation of respiratory rate from photoplethysmographic imaging videos compared to pulse oximetry," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1331–8, 2015.

[194] A. Kikta and P. Augustyniak, "Comparing methods of ECG respiration signals derivation based on measuring the amplitude of QRS complexes," *Journal of Medical Informatics and Technologies*, vol. 11, pp. 155–63, 2007.

[195] V. Kumar and G. Singh, "Estimation of respiration rate from ECG using canonical components analysis and ensemble empirical mode decomposition," *International Journal of Bio-Science and Bio-Technology*, vol. 7, no. 3, pp. 139–46, 2015.

[196] H. H. Kuo *et al.*, "Using ECG surface electrodes in measurement of respiration rate for preterm infants," in *Proc ISBB.* IEEE, 2014, pp. 12–5.

[197] M. M. Lakdawala, "Derivation of the respiratory rate signal from a single lead ECG," MSc. Thesis, New Jersey Institute of Technology, 2008.

[198] J. Lázaro *et al.*, "Deriving respiration from the pulse photoplethysmographic signal," in *Conf Proc CinC.* Hangzhou: IEEE, 2011, pp. 713–6.

[199] J. Lázaro, "Non-invasive techniques for respiratory information extraction based on pulse photoplethysmogram and electrocardiogram," PhD. Thesis, University of Zaragoza, 2015.

[200] J. Lázaro *et al.*, "Respiratory rate derived from smartphone-camera-acquired pulse photoplethysmographic signals," *Physiological Measurement*, vol. 36, no. 11, pp. 2317–33, 2015.

[201] J. Lazaro *et al.*, "Smartphone-camera-acquired pulse photoplethysmographic signal for deriving respiratory rate," in *Conf Proc ESGCO.* IEEE, 2014, pp. 121–2.

[202] E. M. Lee *et al.*, "Respiratory rate detection algorithms by photoplethysmography signal processing," in *Conf Proc Eng Med Biol Soc.* IEEE, 2008, pp. 1140–3.

[203] J. Lee and K. H. Chon, "Respiratory rate extraction via an autoregressive model using the optimal parameter search criterion," *Annals of Biomedical Engineering*, vol. 38, no. 10, pp. 3218–25, 2010.

[204] J. Lee and K. H. Chon, "Time-varying autoregressive model-based multiple modes particle filtering algorithm for respiratory rate extraction from pulse oximeter," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 790–4, 2011.

[205] J. Lee and K. H. Chon, "An autoregressive model-based particle filtering algorithms for extraction of respiratory rates as high as 90 breaths per minute from pulse oximeter," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 9, pp. 2158–67, 2010.

[206] J. Lee, J. P. Florian, and K. H. Chon, "Respiratory rate extraction from pulse oximeter and electrocardiographic recordings," *Physiological Measurement*, vol. 32, no. 11, pp. 1763–73, 2011.

[207] M. Leier, G. Jervan, and W. Stork, "Respiration signal extraction from photoplethysmogram using pulse wave amplitude variation," in *Proc ICC.* IEEE, 2014, pp. 3535–40.

[208] P. Leonard *et al.*, "Standard pulse oximeters can be used to monitor respiratory rate," *Emergency Medicine Journal*, vol. 20, no. 6, pp. 524–5, 2003.

[209] P. a. Leonard *et al.*, "An automated algorithm for determining respiratory rate by photoplethysmogram in children," *Acta Paediatrica*, vol. 95, no. 9, pp. 1124–8, 2006.

[210] P. a. Leonard *et al.*, "A fully automated algorithm for the determination of respiratory rate from the photoplethysmogram," *Journal of Clinical Monitoring and Computing*, vol. 20, no. 1, pp. 33–6, 2006.

[211] P. Leonard *et al.*, "An algorithm for the detection of individual breaths from the pulse oximeter waveform," *Journal of Clinical Monitoring and Computing*, vol. 18, no. 5-6, pp. 309–12, 2004.

[212] D. Li, H. Zhao, and S. Dou, "A new signal decomposition to estimate breathing rate and heart rate from photoplethysmography signal," *Biomedical Signal Processing and Control*, vol. 19, pp. 89–95, 2015.

[213] S. L. Lin, C. K. Chen, and C. S. Chien, "Implementation and validation of ECG-derived respiration with QRS characteristics," *Applied Mechanics and Materials*, vol. 479-80, pp. 457–62, 2013.

[214] L. G. Lindberg, H. Ugnell, and P. Å. Öberg, "Monitoring of respiratory and heart rates using a fibre-optic sensor," *Medical & Biological Engineering & Computing*, vol. 30, no. 5, pp. 533–7, 1992.

[215] G. Liu *et al.*, "Automatic detection of respiratory rate from electrocardiogram, respiration induced plethysmography and 3D acceleration signals," *Journal of Central South University*, pp. 2423–31, 2013.

[216] K. V. Madhav *et al.*, "Use of multi scale PCA for extraction of respiratory activity from photoplethysmographic signals," in *Conf Proc Instrumentation and Measurement Technology.* IEEE, 2012, pp. 1784–7.

[217] K. V. Madhav *et al.*, "Extraction of respiratory activity from ECG and PPG signals using vector autoregressive model," in *Proc Int Symp Medical Measurements and Applications.* IEEE, 2012, pp. 1–4.

[218] K. V. Madhav *et al.*, "Extraction of respiration rate from ECG and BP signals using order reduced-modified covariance AR technique," in *Proc CISP.*, vol. 9. IEEE, 2010, pp. 4059–63.

[219] K. V. Madhav *et al.*, "Robust extraction of respiratory activity from PPG signals using modified MSPCA," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1094–106, 2013.

[220] K. V. Madhav *et al.*, "Estimation of respiration rate from ECG, BP and PPG signals using empirical mode decomposition," in *Conf Proc Instrumentation and Measurement Technology.* IEEE, 2011.

[221] K. V. Madhav *et al.*, "Extraction of respiratory activity from PPG and BP signals using Principal Component Analysis," in *Conf Proc Communications and Signal Processing.* IEEE, 2011, pp. 452–6.

[222] K. V. Madhav *et al.*, "Estimation of respiratory rate from principal components of photoplethysmographic signals," in *Conf Proc IECBES.* IEEE, 2010, pp. 311–4.

[223] K. V. Madhav *et al.*, "A robust signal processing method for extraction of respiratory activity from artifact corrupted PPG signal," in *Proc RAICS.* IEEE, 2011, pp. 451–6.

[224] C. Mason and L. Tarassenko, "Quantitative assessment of respiratory derivation algorithms," in *Conf Proc Eng Med Biol Soc.* IEEE, 2001, pp. 1998–2001.

[225] B. Mazzanti, C. Lamberti, and J. de Bie, "Validation of an ECG-derived respiration monitoring method," in *Conf Proc CinC.* IEEE, 2003, pp. 613–6.

[226] O. Meste, G. Blain, and S. Bermon, "Analysis of the respiratory and cardiac systems coupling in pyramidal exercise using a time-varying model," in *Conf Proc CinC.* IEEE, 2002, pp. 429–32.

[227] L. Mirmohamadsadeghi and J.-M. Vesin, "Real-time multi-signal frequency tracking with a bank of notch filters to estimate the respiratory rate from the ECG," *Physiological Measurement*, vol. 37, no. 9, pp. 1573–87, 2016.

[228] L. Mirmohamadsadeghi and J.-M. Vesin, "Respiratory rate estimation from the ECG using an instantaneous frequency tracking algorithm," *Biomedical Signal Processing and Control*, vol. 14, pp. 66–72, 2014.

[229] L. Mirmohamadsadeghi and J.-m. Vesin, "Estimating the real-time respiratory rate from the ECG with a bank of notch filters," in *Conf Proc CinC.* IEEE, 2015, pp. 581–4.

[230] M. Momot *et al.*, "Estimation of respiratory rate based on ECG signal using regression coefficients and spectral analysis," in *Proc. 14th Int Conf Biomed Eng*, Kaunas, Lithuania, 2010, pp. 49–52.

[231] K. Nakajima, T. Tamura, and H. Miike, "Monitoring of heart and respiratory rates by photoplethysmography using a digital filtering technique," *Medical Engineering & Physics*, vol. 18, no. 5, pp. 365–72, 1996.

[232] K. Nakajima *et al.*, "Photoplethysmographic measurement of heart and respiratory rates using digital filters," in *Conf Proc Eng Med Biol Soc.*, vol. 18, no. 5. IEEE, 1993, pp. 1006–7.

[233] A. Nayan, N. Risman, and R. Jaafar, "Breathing rate estimation from a single-lead electrocardiogram acquisition system," *International Journal of Applied Engineering Research*, vol. 10, no. 17, pp. 38 154–8, 2015.

[234] N. A. Nayan, N. S. Risman, and R. Jaafar, "A portable respiratory rate estimation system with a passive single-lead electrocardiogram acquisition module," *Technology and Health Care*, vol. 24, pp. 591–7, 2016.

[235] L. Nilsson, A. Johansson, and S. Kalman, "Monitoring of respiratory rate in postoperative care using a new photoplethysmographic technique," *Journal of Clinical Monitoring and Computing*, vol. 16, no. 4, pp. 309–15, 2000.

[236] M. Orini *et al.*, "Estimation of spontaneous respiratory rate from photoplethysmography by cross time-frequency analysis," in *Conf Proc CinC.* Hangzhou: IEEE, 2011, pp. 661–4.

[237] C. Orphanidou *et al.*, "Data fusion for estimating respiratory rate from a single-lead ECG," *Biomedical Signal Processing and Control*, vol. 8, no. 1, pp. 98–105, 2013.

[238] C. Orphanidou *et al.*, "Spectral fusion for estimating respiratory rate from the ECG," in *Conf Proc ITAB.* IEEE, 2009, pp. 1–4.

[239] C. Park and B. Lee, "Real-time estimation of respiratory rate from a photoplethysmogram using an adaptive lattice notch filter," *Biomedical Engineering Online*, vol. 13, p. 170, 2014.

[240] S.-B. Park *et al.*, "An improved algorithm for respiration signal extraction from electrocardiogram measured by conductive textile electrodes using instantaneous frequency estimation," *Medical & Biological Engineering & Computing*, vol. 46, no. 2, pp. 147–58, 2008.

[241] M. D. Peláez-Coca *et al.*, "Cross time-frequency analysis for combining information of several sources: Application to estimation of spontaneous respiratory rate from photoplethysmography," *Computational and Mathematical Methods in Medicine*, vol. 2013, p. 631978, 2013.

[242] M. A. F. Pimentel *et al.*, "Probabilistic estimation of respiratory rate using Gaussian processes," in *Conf Proc Eng Med Biol Soc.* IEEE, 2013, pp. 2902–5.

[243] B. Prathyusha, T. Rao, and D. Asha, "Extraction of respiratory rate from PPG signals using PCA and EMD," *International Journal of Research in Engineering and Technology*, vol. 1, no. 2, pp. 164–84, 2012.

[244] K. Rajkumar and K. Ramya, "Respiration rate diagnosis using single lead ECG in real time," *Global Journal of Medical Research*, vol. 13, no. 1, pp. 7–11, 2013.

[245] M. Sakai *et al.*, "Development of lead system for ECG-derived respiration aimed at detection of obstructive sleep apnea syndrome," in *Conf Proc SITIS.* IEEE, 2013, pp. 971–5.

[246] J. Salinger *et al.*, "Measurement of breathing frequency from ECG in the examination of autonomous nervous system activities: suggested methods and their verification," *Acta Univ. Palacki. Olomuc.*, vol. 35, no. 2, pp. 95–103, 2005.

[247] S. Sarkar, S. Bhattacherjee, and S. Pal, "Extraction of respiration signal from ECG for respiratory rate estimation," in *Michael Faraday IET International Summit 2015.* Kolkata: Institution of Engineering and Technology, 2015, pp. 336–40.

[248] N. Selvaraj and K. H. Chon, "Algorithms for real-time detection of motion artifacts and accurate estimation of respiratory rates using pulse oximeter."

[249] R. Sethi *et al.*, "Comparison of respiration rate derived from pulse oximetry and transthoracic impedance," *European Journal of Anaesthesiology*, vol. 31, pp. 30–1, 2014.

[250] S. A. Shah *et al.*, "Continuous measurement of respiration rate using the photoplethysmogram and the electrocardiogram," in *Proc UK & RI Postgraduate Conference in Biomedical Engineering and Medical Physics.*, 2009, pp. 11–2.

[251] H. Sharma, K. Sharma, and O. L. Bhagat, "Respiratory rate extraction from single-lead ECG using homomorphic filtering," *Computers in Biology and Medicine*, vol. 59, pp. 80–6, 2015.

[252] A. Shayei, S. P. Ehsani, and M. Shabany, "Efficient implementation of real-time ECG derived respiration system using cubic spline interpolation," in *Proc ISCAS.*, 2013, pp. 1083–6.

[253] K. H. Shelley *et al.*, "The use of joint time frequency analysis to quantify the effect of ventilation on the pulse oximeter waveform," *Journal of Clinical Monitoring and Computing*, vol. 20, no. 2, pp. 81–7, 2006.

[254] E. F. Sierra-Alonso, L. M. Sep, and R. Bail, "Estimating respiratory frequency from HRV during treadmill exercise testing," in *Conf Proc CinC.* IEEE, 2013, pp. 121–4.

[255] R. Singhathip *et al.*, "Extracting respiration rate from ECG raw signals," *Biomedical Engineering: Applications, Basis and Communications*, vol. 22, no. 04, pp. 307–314, 2010.

[256] A. Sobron, I. Romero, and T. Lopetegi, "Evaluation of methods for estimation of respiratory frequency from the ECG," in *Conf Proc CinC.* IEEE, 2010, pp. 513–6.

[257] L. Sohrt-Petersen, "Evaluation of algorithms for ECG derived respiration in the context of heart rate variability studies," Master's Thesis, Aalborg University, 2013.

[258] K. T. Sweeney *et al.*, "Employing ensemble empirical mode decomposition for artifact removal: extracting accurate respiration rates from ECG data during ambulatory activity," in *Conf Proc Eng Med Biol Soc.* IEEE, 2013, pp. 977–80.

[259] J. Thayer *et al.*, "Estimating respiratory frequency from autoregressive spectral analysis of heart period," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 4, pp. 41–5, 2002.

[260] H. R. W. Touw *et al.*, "Photoplethysmography respiratory rate monitoring in patients receiving procedural sedation and analgesia for upper gastrointestinal endoscopy," *Journal of Clinical Monitoring and Computing [in press]*, 2016.

[261] M. Varanini *et al.*, "Adaptive filtering of ECG signal for deriving respiratory activity," in *Conf Proc CinC.* Chicago, IL: IEEE, 1991, pp. 621–4.

[262] K. V. Madhav *et al.*, "A model based method for deriving respiratory activity from photoplethysmographic signals," in *Conf Proc ISSPA.* IEEE, 2010, pp. 312–5.

[263] N. Vinh *et al.*, "Comparison of two methods for demodulation of pulse signals - application in case of central sleep apnea," *Journal of Science and Technology*, vol. 49, no. 1, 2011.

[264] C. Wang, Z. Li, and X. Wei, "Monitoring heart and respiratory rates at radial artery based on PPG," *Optik - International Journal for Light and Electron Optics*, vol. 124, no. 19, pp. 3954–6, 2013.

[265] S. M. Wendelken *et al.*, "Monitoring respiration rate in PACU patients using the plethysmogram from a pulse oximeter," in *Conf Proc Soc Technology in Anesthesia.*, vol. 3, no. 4, 2005, p. 89.

[266] D. Wertheim *et al.*, "Extracting respiratory data from pulse oximeter plethysmogram traces in newborn infants," *Archives of Disease in Childhood. Fetal and Neonatal Edition*, vol. 94, no. 4, pp. F301–3, 2009.

[267] D. Wertheim *et al.*, "Monitoring respiration in wheezy preschool children by pulse oximetry plethysmogram analysis," *Medical & Biological Engineering & Computing*, vol. 51, no. 9, pp. 965–70, 2013.

[268] D. Widjaja *et al.*, "ECG-derived respiration: comparison and new measures for respiratory variability," in *Conf Proc CinC.* IEEE, 2010, pp. 149–52.

[269] D. Wu, M. Y. M. Wong, and Y.-t. Zhang, "The accuracy of respiratory rate estimation using electrocardiography and photoplethysmography," in *Conf Proc ITAB.* IEEE, 2010.

[270] D. Wu *et al.*, "Automatic estimation of respiratory rate from pulse transit time in normal subjects at rest," in *Conf Proc Eng Med Biol Soc BHI.*, vol. 25. IEEE, 2012, pp. 779–81.

[271] B. Yang and K. H. Chon, "A novel approach to monitor nonstationary dynamics in physiological signals: application to blood pressure, pulse oximeter, and respiratory data," *Annals of Biomedical Engineering*, vol. 38, no. 11, pp. 3478–88, 2010.

[272] W. Yi and K. Park, "Derivation of respiration from ECG measured without subject's awareness using wavelet transform," in *Conf Proc Eng Med Biol Soc.* IEEE, 2002, pp. 130–1.

[273] C. Yi-Hsin, W. Hau-Tieng, and H. Shu-Shya, "ECG-derived respiration and instantaneous frequency based on the synchrosqueezing transform: application to patients with atrial fibrillation," *arXiv preprint*, 2011.

[274] Y. Yoshida, K. Yokoyama, and N. Ishii, "Real-time continuous estimation of respiratory frequency during sleep based on heart rate time series," in *Conf Proc Eng Med Biol Soc.* IEEE, 2007, pp. 648–51.

[275] T. Yoshimura *et al.*, "An ECG electrode-mounted heart rate, respiratory rhythm, posture and behavior recording system." in *Conf Proc Eng Med Biol Soc.* IEEE, 2004, pp. 2373–4.

[276] N. A. Zainudin *et al.*, "Respiratory rate of photoplethysmogram signal from anaesthetic patients," in *Conf Proc ISCAIE.* IEEE, 2015, pp. 171–4.

[277] X. Zhang and Q. Ding, "Respiratory rate monitoring from the photoplethysmogram via sparse signal reconstruction," *Physiological Measurement*, vol. 37, no. 7, pp. 1105–19, 2016.

[278] T. Zhu *et al.*, "Bayesian fusion of algorithms for the robust estimation of respiratory rate from the photoplethysmogram," in *Conf Proc Eng Med Biol Soc.* IEEE, 2015, pp. 6138–41.

[279] S. Fleming *et al.*, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies," *Lancet*, vol. 377, no. 9770, pp. 1011–8, 2011.

[280] S. A. Shah, "Vital sign monitoring and data fusion for paediatric triage," DPhil. Thesis, University of Oxford, 2012.

[281] S. Fleming, "Measurement and fusion of non-invasive vital signs for routine triage of acute paediatric illness," DPhil. Thesis, University of Oxford, 2010.

[282] P. S. Addison and J. N. Watson, "Secondary transform decoupling of shifted nonstationary signal modulation components: application to photoplethysmography," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 1, pp. 43–57, 2004.

[283] L. Sörnmo and P. Laguna, "Electrocardiogram (ECG) signal processing," in *Wiley Encyclopedia of Biomedical Engineering*, 2006.

[284] M. L. Ahlstrom and W. J. Tompkins, "Digital filters for real-time ECG signal processing using microprocessors." *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 9, pp. 708–13, 1985.

[285] P. S. Hamilton, "A comparison of adaptive and nonadaptive filters for reduction of power line interference in the ECG." *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 105–9, 1996.

[286] C. Li, C. Zheng, and C. Tai, "Detection of ECG characteristic points using wavelet transforms," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 1, pp. 21–8, 1995.

[287] W. Zong, G. Moody, and D. Jiang, "A robust open-source algorithm to detect onset and duration of QRS complexes," in *Proc CinC.* IEEE, 2003, pp. 737–40.

[288] J. P. Martínez *et al.*, "A wavelet-based ECG delineator: evaluation on standard databases." *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 570–81, 2004.

[289] I. R. Legarreta *et al.*, "Continuous wavelet transform modulus maxima analysis of the electrocardiogram: beat characterisation and beat-to-beat measurement," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 3, no. 1, pp. 19–42, 2005.

[290] E. Gil *et al.*, "PTT variability for discrimination of sleep apnea related decreases in the amplitude fluctuations of PPG signal in children." *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1079–88, 2010.

[291] D. Widjaja *et al.*, "Application of kernel principal component analysis for single-lead-ECG-derived respiration," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 1169–76, 2012.

[292] R. D. Berger *et al.*, "An efficient algorithm for spectral analysis of heart rate variability," *IEEE Transactions on Biomedical Engineering*, vol. 33, no. 9, pp. 900–4, 1986.

[293] R. Bailón *et al.*, "The integral pulse frequency modulation model with time-varying threshold: application to heart rate variability analysis during exercise stress testing." *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 642–52, 2011.

[294] N. R. Gaddum *et al.*, "Beat-to-beat variation in pulse wave velocity during breathing maneuvers," *Magnetic Resonance in Medicine*, vol. 72, no. 1, pp. 202–10, 2014.

[295] S. A. Shah *et al.*, "Respiratory rate estimation during triage of children in hospitals," *Journal of Medical Engineering & Technology*, vol. 39, no. 8, pp. 514–24, 2015.

[296] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology*, vol. 29, no. 3, pp. 306–10, 1970.

[297] L. Sörnmo and P. Laguna, *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, 1st ed. Academic Press, 2005.

[298] E. Gil, J. María Vergara, and P. Laguna, "Detection of decreases in the amplitude fluctuation of pulse photoplethysmography signal as indication of obstructive sleep apnea syndrome in children," *Biomedical Signal Processing and Control*, vol. 3, no. 3, pp. 267–277, 2008.

[299] M. Saeed *et al.*, "Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database," *Critical Care Medicine*, vol. 39, no. 5, pp. 952–60, 2011.

[300] J. Welch *et al.*, "The Massachusetts General Hospital-Marquette Foundation Hemodynamic and Electrocardiographic Database – Comprehensive collection of critical care waveforms," *Journal of Clinical Monitoring*, vol. 7, no. 1, pp. 96–7, 1991.

[301] G. Moody and R. Mark, "A database to support development and evaluation of intelligent intensive care monitoring," in *Proc CinC.* IEEE, 1996, pp. 657–60.

[302] N. Iyengar *et al.*, "Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics," *The American Journal of Physiology*, vol. 271, no. 4 Pt 2, pp. R1078–84, 1996.

[303] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–20, 2000.

[304] M. A. García-gonzález *et al.*, "A comparison of heartbeat detectors for the seismocardiogram," in *Proc CinC.* IEEE, 2013, pp. 461–4.

[305] Y. Ichimaru and G. B. Moody, "Development of the polysomnographic database on CD-ROM," *Psychiatry and Clinical Neurosciences*, vol. 53, no. 2, pp. 175–7, 1999.

[306] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–66, 2005.

[307] T. Penzel *et al.*, "The apnea-ECG database," in *Proc CinC.* IEEE, 2000, pp. 255–8.

[308] Portland State University Biomedical Signal Processing Lab, "Traumatic brain injury data."

[309] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, pp. 307–10, 1986.

[310] L. Mason, "Signal Processing Methods for Non-Invasive Respiration Monitoring," DPhil. Thesis, University of Oxford, 2002.

[311] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[312] A. Boardman *et al.*, "A study on the optimum order of autoregressive models for heart rate variability," *Physiological Measurement*, vol. 23, no. 2, pp. 325–36, may 2002.

[313] C. Shannon, "Communication in the presence of noise (reprint of classic paper)," *Proc. IEEE*, vol. 86, no. 2, pp. 447–457, 1998.

[314] M. Piepoli *et al.*, "Autonomic control of the heart and peripheral vessels in human septic shock," *Intensive Care Medicine*, vol. 21, no. 2, pp. 112–9, 1995.

[315] R. Pallás-Areny, J. Colominas-Balagué, and F. J. Rosell, "The effect of respiration-induced heart movements on the ECG." *IEEE Transactions on Biomedical Engineering*, vol. 36, no. 6, pp. 585–90, 1989.

[316] R. J. Gibbons *et al.*, "ACC/AHA guidelines for exercise testing: a report of the American College of Cardiology/American Heart Association task force on practice guidelines (Committee on Exercise Testing)," *Journal of the American College of Cardiology*, vol. 30, no. 1, pp. 260–311, 1997.

[317] J. Lian, L. Wang, and D. Muessig, "A simple method to detect atrial fibrillation using RR intervals," *The American Journal of Cardiology*, vol. 107, no. 10, pp. 1494–7, 2011.

[318] L. Chen *et al.*, "Algorithms to qualify respiratory data collected during the transport of trauma patients," *Physiological Measurement*, vol. 27, no. 9, pp. 797–816, 2006.

[319] L. Nilsson *et al.*, "Combined photoplethysmographic monitoring of respiration rate and pulse: a comparison between different measurement sites in spontaneously breathing subjects," *Acta Anaesthesiologica Scandinavica*, vol. 51, no. 9, pp. 1250–7, 2007.

[320] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals," *Current Cardiology Reviews*, vol. 8, no. 1, pp. 14–25, 2012.

[321] J. Alastruey *et al.*, "Pulse wave propagation in a model human arterial network: Assessment of 1-D visco-elastic simulations against in vitro measurements," *Journal of Biomechanics*, vol. 44, no. 12, pp. 2250–8, 2011.

[322] K. H. Shelley *et al.*, "What is the best site for measuring the effect of ventilation on the pulse oximeter waveform?" *Anesthesia and Analgesia*, vol. 103, no. 2, pp. 372–7, 2006.

[323] J. M. Feldman, "Can clinical monitors be used as scientific instruments?" *Anesthesia and Analgesia*, vol. 103, no. 5, pp. 1071–1072, 2006.

[324] K. H. Shelley, "Photoplethysmography: beyond the calculation of arterial oxygen saturation and heart rate," *Anesthesia and Analgesia*, vol. 105, no. 6, pp. S31–6, 2007.

[325] L. Ø. Høiseth *et al.*, "Respiratory variations in the photoplethysmographic waveform amplitude depend on type of pulse oximetry device," *Journal of Clinical Monitoring and Computing*, vol. 30, no. 3, pp. 317–25, 2015.

[326] I. Constant *et al.*, "Pulse rate variability is not a surrogate for heart rate variability," *Clinical Science*, vol. 97, no. 4, pp. 391–397, 1999.

[327] N. Selvaraj *et al.*, "Influence of respiratory rate on the variability of blood volume pulse characteristics," *Journal of Medical Engineering & Technology*, vol. 33, no. 5, pp. 370–5, 2009.

[328] R. Bailón, L. Sörnmo, and P. Laguna, "A robust method for ECG-based estimation of the respiratory frequency during stress testing," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 7, pp. 1273–85, 2006.

[329] Y. Nam, J. Lee, and K. H. Chon, "Respiratory rate estimation from the built-in cameras of smartphones and tablets," *Annals of Biomedical Engineering*, vol. 42, no. 4, pp. 885–98, 2014.

[330] J. Li *et al.*, "Comparison of respiratory-induced variations in photoplethysmographic signals," *Physiological Measurement*, vol. 31, no. 3, pp. 415–25, 2010.

[331] C. W. Seymour *et al.*, "Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3)," *Journal of the American Medical Association*, vol. 315, no. 8, pp. 762–74, 2016.

[332] J. A. Hirsch and B. Bishop, "Respiratory sinus arrhythmia in humans: how breathing pattern modulates heart rate," *The American Journal of Physiology*, vol. 241, no. 4, pp. H620–9, 1981.

[333] J. Lázaro *et al.*, "Respiratory Rate Influence in the Resulting Magnitude of Pulse Photoplethysmogram Derived Respiration Signals," in *Proc CinC.* Cambridge, MA: IEEE, 2014, pp. 289–92.

[334] D. Caggiano and S. Reisman, "Respiration derived from the electrocardiogram: a quantitative comparison of three different methods," in *Conf Proc Northeast Bioengineering.* IEEE, 1996, pp. 103–4.

[335] J. Mateo and P. Laguna, "Analysis of heart rate variability in the presence of ectopic beats using the heart timing signal," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 334–43, 2003.

[336] K. H. Hamed, "Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis," *Journal of Hydrology*, vol. 349, no. 3-4, pp. 350–363, 2008.

[337] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

[338] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[339] Z. Šidák, "Rectangular confidence regions for the means of multivariate normal distributions," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 626–33, 1967.

[340] A. I. Vinik and D. Ziegler, "Diabetic cardiovascular autonomic neuropathy," *Circulation*, vol. 115, no. 3, pp. 387–97, 2007.

[341] M. A. Pimentel *et al.*, "Towards a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering [in press]*, 2016.

[342] N. a. Obuchowski *et al.*, "Quantitative imaging biomarkers: A review of statistical methods for computer algorithm comparisons," *Statistical Methods in Medical Research*, vol. 24, no. 1, pp. 68–106, 2015.

[343] J. M. Bland and D. G. Altman, "Agreement between methods of measurement with multiple observations per individual," *Journal of Biopharmaceutical Statistics*, vol. 17, no. 4, pp. 571–82, 2007.

[344] H. X. Barnhart, M. J. Haber, and L. I. Lin, "An overview on assessing agreement with continuous measurements," *Journal of Biopharmaceutical Statistics*, vol. 17, no. 4, pp. 529–69, 2007.

[345] R. Bailon *et al.*, "Robust electrocardiogram derived respiration from stress test recordings: validation with respiration recordings," in *Conf Proc CinC.* IEEE, 2004, pp. 293–6.

[346] L. Nilsson, A. Johansson, and S. Kalman, "Respiration can be monitored by photoplethysmography with high sensitivity and specificity regardless of anaesthesia and ventilatory mode." *Acta Anaesthesiologica Scandinavica*, vol. 49, no. 8, pp. 1157–62, 2005.

[347] G. Hackmann *et al.*, "Toward a two-tier clinical warning system for hospitalized patients," in *AMIA Annual Symposium Proceedings*, 2011, pp. 511–9.

[348] L. Clifton *et al.*, "Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 722–30, 2014.

[349] M. R. Pinsky, "Hemodynamic evaluation and monitoring in the ICU," *Chest*, vol. 132, no. 6, pp. 2020–9, 2007.

[350] Philips Healthcare, "Wireless Biosensor webpage."

[351] a. Mäkivirta *et al.*, "The median filter as a preprocessor for a patient monitor limit alarm system in intensive care," *Computer Methods and Programs in Biomedicine*, vol. 34, no. 2-3, pp. 139–44, 1991.

[352] G. D. Clifford *et al.*, "Robust parameter extraction for decision support using multimodal intensive care data," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1887, pp. 411–29, 2009.

[353] P. Yang, G. A. Dumont, and J. M. Ansermino, "Sensor fusion using a hybrid median filter for artifact removal in intraoperative heart rate monitoring," *Journal of Clinical Monitoring and Computing*, vol. 23, no. 2, pp. 75–83, 2009.

[354] A. Hann, "Multi-parameter Monitoring for Early Warning of Patient Deterioration," DPhil. Thesis, Unviersity of Oxford, 2008.

[355] S. Khalid *et al.*, "A Two-Class Approach to the Detection of Physiological Deterioration in Patient Vital Signs, with Clinical Label Refinement." *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1231–8, 2012.

[356] D. Wong, "Identifying Vital Sign Abnormality in Acutely-Ill Patients," DPhil. Thesis, University of Oxford, 2011.

[357] L. Tarassenko *et al.*, "Novelty Detection," in *Encyclopedia of Structural Health Monitoring*, C. Boller, F.-K. Chang, and Y. Fujino, Eds. John Wiley & Sons, 2009, ch. 35, pp. 1–23.

[358] D. Arthur and S. Vassilvitskii, "K-Means++: the Advantages of Careful Seeding," in *Proc ACM-SIAM symposium on discrete algorithms.*, vol. 8, 2007, pp. 1027–35.

[359] M. A. F. Pimentel, "Modelling of Vital-Sign Data from Post-operative Patients," DPhil. Thesis, University of Oxford, 2015.

[360] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–76, 1962.

[361] M. Hravnak *et al.*, "Cardiorespiratory instability before and after implementing an integrated monitoring system." *Critical Care Medicine*, vol. 39, no. 1, pp. 65–72, 2011.

[362] C. P. Subbe *et al.*, "Validation of a modified Early Warning Score in medical admissions," *QJM : Monthly Journal of the Association of Physicians*, vol. 94, no. 10, pp. 521–6, 2001.

[363] S. Romero-Brufau *et al.*, "Why the C-statistic is not informative to evaluate early warning scores and what metrics to use," *Critical Care*, vol. 19, no. 1, p. 285, 2015.

[364] W. H. Maisel, J. D. Rawn, and W. G. Stevenson, "Atrial fibrillation after cardiac surgery," *Annals of Internal Medicine*, vol. 135, no. 12, pp. 1061–1073, 2001.

[365] M. M. Churpek *et al.*, "Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards," *Critical Care Medicine*, vol. 44, no. 2, pp. 368–374, 2016.

[366] P. Schulam and S. Saria, "A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure: Supplement," *Advances in Neural Information Processing Systems (NIPS)*, no. 5, pp. 748–756, 2015.

[367] M. M. Churpek, R. Adhikari, and D. P. Edelson, "The value of vital sign trends for detecting clinical deterioration on the wards," *Resuscitation*, vol. 102, pp. 1–5, 2016.

[368] A. Bates *et al.*, "Respiratory rate and flow waveform estimation from tri-axial accelerometer data," in *Conf Proc Body Sensor Networks.* Singapore: IEEE, 2010, pp. 144–50.

[369] J. Mann *et al.*, "Simultaneous Activity and Respiratory Monitoring Using an Accelerometer," *Conf Proc Body Sensor Networks.*, pp. 139–43, 2011.

[370] K. Pandia *et al.*, "Extracting respiratory information from seismocardiogram signals acquired on the chest using a miniature accelerometer," *Physiological Measurement*, vol. 33, no. 10, pp. 1643–60, 2012.

[371] T. G. Buchman, "Novel representation of physiologic states during critical illness and recovery." *Critical Care*, vol. 14, no. 2, p. 127, 2010.

[372] W. W. Melek *et al.*, "Comparison of trend detection algorithms in the analysis of physiological time-series data." *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 4, pp. 639–51, 2005.

[373] T. G. Buchman, P. K. Stein, and B. Goldstein, "Heart rate variability in critical illness and critical care." *Current Opinion in Critical Care*, vol. 8, no. 4, pp. 311–5, 2002.

[374] A. J. E. Seely and P. T. Macklem, "Complex systems and the technology of variability analysis." *Critical Care*, vol. 8, no. 6, pp. R367–84, 2004.

[375] U. Rajendra Acharya *et al.*, "Heart rate variability: a review," *Medical & Biological Engineering & Computing*, vol. 44, no. 12, pp. 1031–51, 2006.

[376] C. A. Buchan, A. Bravi, and A. J. E. Seely, "Variability analysis and the diagnosis, management, and treatment of sepsis." *Current Infectious Disease Reports*, vol. 14, no. 5, pp. 512–21, 2012.

[377] A. Bravi, A. Longtin, and A. J. E. Seely, "Review and classification of variability analysis techniques with clinical applications." *Biomedical Engineering Online*, vol. 10, no. 1, p. 90, 2011.

[378] V. E. Papaioannou *et al.*, "Temperature variability analysis using wavelets and multiscale entropy in patients with systemic inflammatory response syndrome, sepsis and septic shock," *Critical Care*, vol. 16, no. 2, p. R51, 2012.

[379] A. J. E. Seely, G. C. Green, and A. Bravi, "Continuous Multiorgan Variability monitoring in critically ill patients–complexity science at the bedside." in *Conf Proc Eng Med Biol Soc.* IEEE, 2011, pp. 5503–6.

[380] A. Dupre, S. Vincent, and P. A. Iaizzo, "Basic ECG Theory, Recordings, and Interpretation," *Handbook of Cardiac Anatomy, Physiology, and Devices*, pp. 191–201, 2005.

[381] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, pp. R1–39, 2007.

[382] S. C. Millasseau *et al.*, "Contour analysis of the photoplethysmographic pulse measured at the finger," *Journal of Hypertension*, vol. 24, no. 8, pp. 1449–56, 2006.

[383] M. Garbarino *et al.*, "Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *Proc MOBIHEALTH.*, 2015, pp. 39–42.